DATA-DRIVEN PSP LINKAGES FOR ATOMISTIC DATASETS

A Dissertation
Presented to
The Academic Faculty

by

Joshua A. Gomberg

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Materials Science and Engineering

Georgia Institute of Technology
August 2017

DATA-DRIVEN PSP LINKAGES FOR ATOMISTIC DATASETS

Approved by:

Dr. Surya R. Kalidindi, Advisor
School of Mechanical Engineering
Georgia Institute of Technology

Dr. David McDowell
School of Materials Science and Engineering
Georgia Institute of Technology

Dr. Mo Li
School of Materials Science and Engineering
Georgia Institute of Technology

Dr. Ben Haaland
School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Hamid Garmestani
School of Materials Science and Engineering
Georgia Institute of Technology

Date Approved: May 11, 2017

ACKNOWLEDGEMENTS

Joshua Gomberg would like to acknowledge Drs. Surya Kalidindi and Andrew Medford for their guidance. Additionally, Dr. Srikanth Patala and Drs. Chandler Becker and Zachary Trautt contributed simulated data upon which much of this work is based. Work was supported by the National Institute for Standards and Technology (No. 70NANB14H191).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ANOVA	Analysis of Variance
ASCA	ANOVA Single Component Analysis
ATGB	Asymmetric Tilt Grain Boundary
CS	Centro-Symmetry
CSL	Coincidence Site Lattice
FEM	Finite Element Method
GB	Grain Boundary
HPC	High-Performance Computing
KDE	Kernel Density Estimation
MD	Molecular Dynamics
MKS	Materials Knowledge Systems
MM	Molecular Mechanics
PC	Principal Component
PCA	Principal Component Analysis
PCF	Pair Correlation Function
PDF	Probability Distribution Function
PSP	Process-Structure-Property
STGB	Symmetric Tilt Grain Boundary
SVD	Singular Value Decomposition

SUMMARY

Multiscale modeling provides a class of methods that allow the behavior of materials to be characterized using empirical and theoretical models across many length scales. This is accomplished through the construction of data-driven learning models that distill knowledge pertaining to the relationships among a material's processing, structure, and properties (the PSP paradigm) from these length-scale dependent models. Much of the recent progress in the field of multiscale modeling has been focused on analysis of mesoscale datasets, where structures are characterized by material composition in discretized spatial regions. In these models, the structure is typically quantified using descriptors such as the pair correlation function or the 2-point spatial correlations (also called the 2-point statistics); an ensemble of these descriptors is then typically represented in a low rank form.

In this work, the multiscale modeling framework is adapted and extended to apply to datasets derived from atomistic simulations, with a primary focus on molecular mechanics (energy minimized structures) and molecular dynamics (evolution of structure with time). In these datasets, structure is described by a list of atomic positions in continuous space, which can be classified as point-cloud data. For datasets such as these, a method for calculating the discretized 2-point statistics is devised that is independent of the simulation box size, which can fluctuate over the course of a molecular dynamics simulation. The efficacy of this method is demonstrated in a study where interatomic potentials of aluminum are categorized by their resulting simulated structures. For the case of grain boundary simulations, an algorithm is described for identifying and

characterizing the structure of atoms that lie within the grain boundary. This algorithm is implemented in additional analytical studies on two datasets: a set of simulations of asymmetric and symmetric tilt grain boundaries where dimensionality reduction is respectively achieved using principal component analysis (PCA) and ANOVA single component analysis (ASCA), an extension of PCA where prior knowledge is used to separately evaluate the covariance structure for different sources of variance within the dataset.

CHAPTER 1. INTRODUCTION

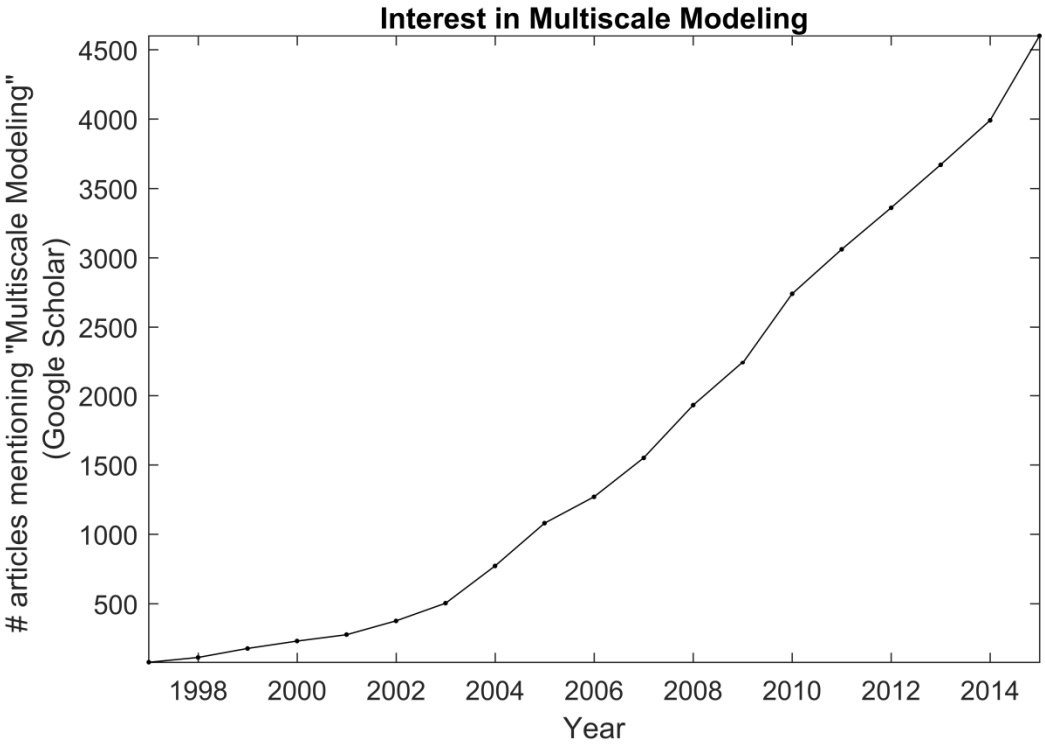

Figure 2.1.1 – Research interest in multiscale modeling.

Multiscale modeling [1-4] gives us a class of techniques that address material properties ranging from the atomic scale to the continuum. When it comes time to manufacture new materials, more detailed structure knowledge allows for far less trial and error, which allows new materials to be synthesized quicker and more cheaply.[5] Currently, there are three primary areas of concern that hinder widespread application of multiscale modeling[6]: the availability of accurate and reliable models (Model Maturity), the seamless integrations of models covering multiple length scales (Model Interoperability), and the ability to tailor processes in materials synthesis to yield

specifically targeted properties (Model Inversion). Interest in multiscale models has been steadily increasing for nearly 20 years now, as evidenced by Figure 2.1.1.

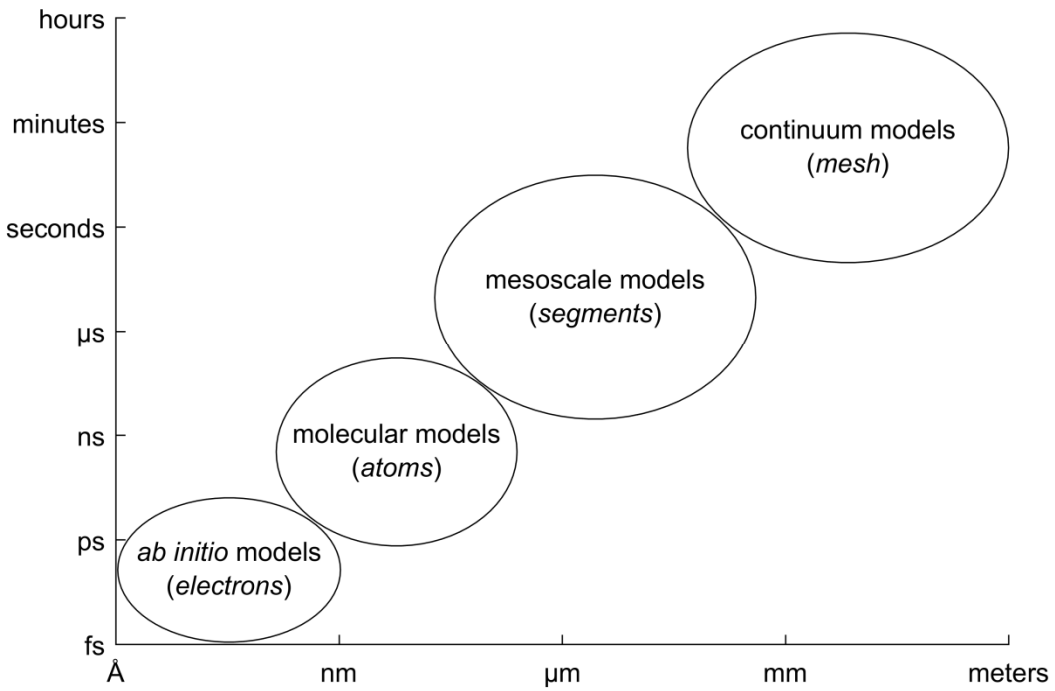

Figure 2.1.2 – Time- and length-scales of simulation techniques.

Remarkable progress in the field of physics-driven models has made possible the ability to numerically simulate a broad range of materials phenomena [1-4, 7-17]. However, it should be noted that modeling techniques governed by physics and are only valid over a limited range of time- and length-scales.[18] Figure 2.1.2 above illustrates the relevant scales for different simulation techniques. *Ab initio* modeling techniques such as Hartree-Fock or density functional theory are based in quantum mechanics and governed by the physics of electrons.[19] Molecular dynamics and mechanics (MM/MD) focus on interactions between atoms[20], and mesoscale models such as dissipative particle dynamics[21] are governed between by interactions between segments or groups

of atoms. At the continuum level, the finite element method (FEM)[22] solves for boundary values of partial differential equations defined on a mesh. It is clear that these classes of simulations operate under different sets of assumptions, which is the reason for the varying scales of accuracy.

A multiscale modeling approach would use simulations on a smaller scale to inform, in some way, the structure and properties of a model at a larger scale. This would get around the limitations imposed by the assumptions made by each class of simulations. However, for such an approach to be effectively utilized, the linkages between simulations of different scales should capture as much of the pertinent knowledge as possible while still being substantially less computationally expensive than repeatedly re-running the smaller scale simulations that inform them.

An approach such as this requires the development of methods to learn from simulations. For the purpose of multiscale modeling of materials, two different types of learning models are relevant, which operate under the “process-structure-property” (PSP) paradigm. A process-structure model links the conditions under which a material is made with some structure-derived metric. A structure-property model connects the structure with a material property of interest. Models such as this are not concerned with the physics of the underlying simulations. Instead, these linkages are constructed using methods informed by data science.[23-26] In other words, it is the data itself and not the simulation physics that identifies the important features when linking process to structure, or structure to property. Interest in data science is primarily concentrated on distilling high value information from all available data, generated by either simulations or experiments. This emerging cross-disciplinary field is being built on the foundations of

applied mathematics, systems theory, and computational and statistical sciences. Since physics is not a factor, similar methods for constructing learning models at one length scale should also be applicable for learning at a different length scale. Substantial focus has been placed on developing learning models for materials at the mesoscale.[27-33] In the proposed work, these methods are adapted and extended to the regime of atomistic simulation data, with a particular focus on simulations of grain boundaries (GBs).

For a variety of materials, atomic-scale modeling techniques such as MM/MD are commonly employed as a means of investigating fundamental properties, including both structural and chemical responses.[34, 35] In general, MD simulations run on high performance computing (HPC) infrastructures can yield vast amounts of pertinent data for a wide range of structures and simulation conditions. For example, an investigation of GB motion among 388 simulated nickel GBs identified 15 unique trends [36]. From a materials science perspective, the computations underpinning MM/MD simulations can be cast as highly complex "process-structure" relationships. In the example of GB simulations, the "process" variables would describe the methods that control the evolution of the structure, such as the thermodynamic ensemble, force field and applied loads, as well as the configurational constraints governing the initial structure, such as the macro degrees of freedom.. The "structure" would correspond to the elements, configuration and bonding structure of the atoms in a given composition. The concept of a "process-structure" relationship for these atomistic simulations would establish a quantitative connection between the process inputs of the simulation and the resulting atomic-scale structure (output). There has not yet been a systematic effort focused on the extraction of reduced-order "process-structure" linkages capable of rapidly predicting

atomic structures as a function of simulation inputs. This class of functions with their exceptionally low computational cost offer a unique practical approach for addressing inverse problems where one seeks to identify the process recipes that are likely to result in a desired atomic structure.

Another important type of knowledge produced from molecular dynamics/mechanics simulations can be described by linkages between atomic-scale structure and a relevant property such as the overall system energy; these linkages may be categorized as "structure-property" relationships. In the case of GB simulations, GB energies play a vital role in the multiscale modeling of materials phenomena, as they serve as a key input to simulations at a larger scale (e.g., plasticity, failure, recrystallization[34]). While force-field based calculations are significantly less computationally expensive than their quantum-mechanical counterparts, the datasets often investigated are large in size (10^3 – 10^9 atoms) and high-dimensional, and thus cumbersome for use in multi-scale models[34, 37]. Some progress has been made in training neural network potentials to results of quantum mechanical methods such as density functional theory for use in molecular dynamics simulations[38-40] but these methods typically require extremely large training sets (10^3-10^4 systems). Data-science techniques have also been previously applied for the systematic analysis and knowledge extraction from large MM/MD datasets[41-45] with a focus primarily on proteins and other large biomolecules. Within the materials science community, there has been relatively little effort devoted to a systematic analysis and dimensional reduction of force-field based simulations. This is of particular importance given the recent rise in multiscale and hierarchical methods. [29, 46]

It is the goal of this work to adapt the methods for deriving data-driven PSP linkages at the mesoscale to datasets generated by atomistic simulation. As the structures represented in these datasets consist of atomic positions in continuous space (also called "point cloud" datasets), the framework as originally devised for discretized mesoscale structures must be modified to address the challenges posed by the nature of the data. The first issue to be addressed is the calculation of digitized 2-point statistics for point-cloud datasets. This is accomplished by exploiting the properties of convolutions and cross-correlations, as described in Section 2.1. Focusing on the case of GB simulations, the next issue to be addressed is the differentiation of GB atoms from atoms in the bulk crystal, as well as an appropriate structure characterization, which is discussed in Section 2.2. In the case where there exists some prior knowledge of the source of variance within a dataset, which is frequently the case for simulated data, dimensionality reduction may best be achieved by ANOVA single-component analysis (ASCA)[47, 48], an extension of principal component analysis (PCA)[49]. The framework for implementing ASCA, as well as the implementation of ASCA and PCA in linear regression models, is discussed in Section 2.3.

The efficacy of these methods is demonstrated in three analytical studies. First, the utility of PCA analysis of 2-point statistics of atomic structures generated by MD is illustrated in a study where different interatomic potentials are categorized based on simulations performed at different temperatures (CHAPTER 3, APPENDIX A). Next, process-structure and structure-property models are constructed for a dataset of aluminum asymmetric tilt GBs (ATGBs) simulated with MM (Section CHAPTER 4, APPENDIX

B). Lastly, these models are further refined by the implementation of ASCA for a dataset of MM-simulated symmetric tilt GBs (STGBs) of aluminum (Section CHAPTER 5).

CHAPTER 2. EXTENSIONS TO EXISTING FRAMEWORK

2.1 2-Point Statistics for Point Cloud Datasets

2.1.1 Convolution and Cross-Correlation

At a high level, the 2-point statistics function represents the probability of finding the structure in state 1 in one location of a material and the structure in state 2 in another location separated by some displacement vector $\vec{r}$. At the mesoscale, the states 1 and 2 typically represent different phases of the material. At the atomic scale, most of the volume comprising the material consists of empty space. For these sorts of data sets, the states represent if a given location is occupied by vacuum or by an atom of a particular element. The 2-point statistics of simulated atomic structures in continuous space can be determined explicitly by exploiting the properties of convolution and cross-correlation. The convolution and cross-correlation of functions $\mathrm{f}(t)$ and $\mathrm{g}(t)$ are defined, respectively, for $t \in \mathbb{R}^n$ as[50]:

$$(\mathrm{f} * \mathrm{g})(t) = \int_{\mathbb{R}^n} \mathrm{f}(\tau)\mathrm{g}(t-\tau)d\tau \tag{1}$$

$$(\mathrm{f} \star \mathrm{g})(t) = \int_{\mathbb{R}^n} \bar{\mathrm{f}}(\tau)\mathrm{g}(t+\tau)d\tau \tag{2}$$

where $\bar{\mathrm{f}}(\tau)$ is the complex conjugate of $\mathrm{f}(\tau)$. A comprehensive overview of the properties of convolutions is beyond the scope of this text. However, there are a few properties that are

important to highlight for reasons that will be examined later. Convolutions possess both the properties of commutativity and associativity, as demonstrated in Equations 3 and 4 below:

$$(\mathrm{f} * \mathrm{g})(t) = (\mathrm{g} * \mathrm{f})(t) \tag{3}$$

$$(\mathrm{f} * \mathrm{g})(t) * \mathrm{h}(t) = \mathrm{f}(t) * (\mathrm{g} * \mathrm{h})(t) \tag{4}$$

Also, the complex conjugate of the convolution of two functions is equal to the convolution of the complex conjugate of two functions:

$$\overline{(\mathrm{f} * \mathrm{g})}(t) = \bar{\mathrm{f}}(t) * \bar{\mathrm{g}}(t) \tag{5}$$

Additionally, it is useful to note that a cross-correlation can be expressed in terms of a convolution:

$$(\mathrm{f} \star \mathrm{g})(t) = \bar{\mathrm{f}}(-t) * \mathrm{g}(t) \tag{6}$$

2.1.2 Atomic Positions as Real-Space Functions

The structure of a material resulting from an atomic simulation consists of a series of coordinates in continuous space. Mathematically, this structure may be represented as a sum of delta functions, defined such that $\delta(0) = 1$ and $\delta(z) = 0$ for $z \neq 0$. If $\mathcal{S}$ is the set of coordinates in $\mathbb{R}^3$corresponding to atom centers of a single element, then the structure may be represent0ed as

$$\iota_{\mathrm{C}}(\vec{r}|\mathcal{S}) = \sum_{\vec{\tau}\in\mathcal{S}} \delta(\|\vec{\tau} - \vec{r}\|) \tag{7}$$

where $\vec{r} \in \mathbb{R}^3$ and $\|\vec{z}\|$ is the Euclidean norm of vector $\vec{z}$. For any practical visualization of atomic structure, some finite volume must be assigned to each atom. Perhaps the most straightforward approach would be a hard sphere model of uniform density. A uniformly dense sphere of radius R may be expressed in terms of a Heaviside step function $\mathrm{H}(z)$ as

$$\iota_{\mathrm{A}}(\vec{r}|R) = \mathrm{H}(R - \|\vec{r}\|) \tag{8}$$

The continuous mathematical representation of the structure corresponding to a single element, where the atoms are represented by spheres, can be expressed as a convolution of the two functions described in Equations 7 and 8:

$$\iota_{\mathrm{F}}(\vec{r}|R,\mathcal{S}) = \iota_{\mathrm{C}}(\vec{r}|\mathcal{S}) * \iota_{\mathrm{A}}(\vec{r}|R) = \sum_{\vec{\tau}\in\mathcal{S}} \mathrm{H}(R - \|\vec{\tau} - \vec{r}\|) \tag{9}$$

2.1.3 Cross-Correlation of Atomic Position Functions

The unnormalized 2-point statistics function for atoms corresponding to elements labeled 1 and 2 (which may be the same element) can be expressed in terms of a cross-correlation:

$$\alpha_{\mathrm{F}}^{1,2}(\vec{r}) = \iota_{\mathrm{F}}(\vec{r}|R_1,\mathcal{S}_1) \star \iota_{\mathrm{F}}(\vec{r}|R_2,\mathcal{S}_2) \tag{10}$$

where $\mathcal{S}_1$ and $\mathcal{S}_2$ are the sets of atomic coordinates and R_1 and R_2 are the atomic radii corresponding to elements 1 and 2, respectively. Substituting Equation 9 into Equation 10 yields:

$$\alpha_{\mathrm{F}}^{1,2}(\vec{r}) = [\iota_{\mathrm{C}}(\vec{r}|\mathcal{S}_1) * \iota_{\mathrm{A}}(\vec{r}|R_1)] \star [\iota_{\mathrm{C}}(\vec{r}|\mathcal{S}_2) * \iota_{\mathrm{A}}(\vec{r}|R_2)] \tag{11}$$

Employing the properties described in Equations 5 and 6 allows $\alpha_{\mathrm{F}}^{1,2}$ to be represented strictly in terms of convolutions:

$$\alpha_{\mathrm{F}}^{1,2}(\vec{r}) = \left[\overline{\iota_{\mathrm{C}}(-\vec{r}|\mathcal{S}_1)} * \overline{\iota_{\mathrm{A}}(-\vec{r}|R_1)}\right] * [\iota_{\mathrm{C}}(\vec{r}|\mathcal{S}_2) * \iota_{\mathrm{A}}(\vec{r}|R_2)] \tag{12}$$

Rearranging the parenthetical groups as permitted by Equation 4 yields:

$$\alpha_{\mathrm{F}}^{1,2}(\vec{r}) = \overline{\iota_{\mathrm{C}}(-\vec{r}|\mathcal{S}_1)} * \left[\overline{\iota_{\mathrm{A}}(-\vec{r}|R_1)} * \iota_{\mathrm{C}}(\vec{r}|\mathcal{S}_2)\right] * \iota_{\mathrm{A}}(\vec{r}|R_2) \tag{13}$$

The commutative property demonstrated in Equation 3 allows for the reordering of the terms in parenthesis:

$$\alpha_{\mathrm{F}}^{1,2}(\vec{r}) = \overline{\iota_{\mathrm{C}}(-\vec{r}|\mathcal{S}_1)} * \left[\iota_{\mathrm{C}}(\vec{r}|\mathcal{S}_2) * \overline{\iota_{\mathrm{A}}(-\vec{r}|R_1)}\right] * \iota_{\mathrm{A}}(\vec{r}|R_2) \tag{14}$$

Reapplying the associative property described in Equation 4 produces the following result:

$$\alpha_{\mathrm{F}}^{1,2}(\vec{r}) = \left[\overline{\iota_{\mathrm{C}}(-\vec{r}|\mathcal{S}_1)} * \iota_{\mathrm{C}}(\vec{r}|\mathcal{S}_2)\right] * \left[\overline{\iota_{\mathrm{A}}(-\vec{r}|R_1)} * \iota_{\mathrm{A}}(\vec{r}|R_2)\right] \tag{15}$$

Finally, reapplying Equation 6 allows us to express $\alpha_{\mathrm{F}}^{1,2}$ in terms of the convolution of two cross-correlations:

$$\alpha_{\mathrm{F}}^{1,2}(\vec{r}) = \left(\alpha_{\mathrm{C}}^{1,2} * \alpha_{\mathrm{A}}^{1,2}\right)(\vec{r}) \tag{16}$$

Where the functions $\alpha_{\mathrm{C}}^{1,2}$ and $\alpha_{\mathrm{A}}^{1,2}$ are defined, respectively, as:

$$\alpha_{\mathrm{C}}^{1,2}(\vec{r}) = \iota_{\mathrm{C}}(\vec{r}|\mathcal{S}_1) \star \iota_{\mathrm{C}}(\vec{r}|\mathcal{S}_2) \tag{17}$$

$$\alpha_{\mathrm{A}}^{1,2}(\vec{r}) = \iota_{\mathrm{A}}(\vec{r}|R_1) \star \iota_{\mathrm{A}}(\vec{r}|R_2) \tag{18}$$

$\alpha_{\mathrm{C}}^{1,2}$ represents the cross-correlation of the atomic centers. Using Equation 2, we can express this as:

$$\alpha_{\mathrm{C}}^{1,2}(\vec{r}) = \int_{\mathbb{R}^3} \iota_{\mathrm{C}}(\vec{\tau}|\mathcal{S}_1)\, \iota_{\mathrm{C}}(\vec{r}+\vec{\tau}|\mathcal{S}_2) d\vec{\tau} \tag{19}$$

Substituting Equation 7 into Equation 19 yields the expression for $\alpha_{\mathrm{C}}^{1,2}$:

$$\alpha_{\mathrm{C}}^{1,2}(\vec{r}) = \sum_{\vec{\tau}_1 \in \mathcal{S}_1} \sum_{\vec{\tau}_2 \in \mathcal{S}_2} \delta(\|\vec{\tau}_2 - \vec{\tau}_1 - \vec{r}\|) \tag{20}$$

$\alpha_{\mathrm{A}}^{1,2}$ represents the cross-correlation of two spheres of radius R_1 and R_2:

$$\alpha_{\mathrm{A}}^{1,2}(\vec{r}) = \int_{\mathbb{R}^3} \iota_{\mathrm{A}}(\vec{\tau}|R_1)\, \iota_{\mathrm{A}}(\vec{r}+\vec{\tau}|R_2) d\vec{\tau} \tag{21}$$

In Equations 19 and 21, ι_{C} and ι_{A} are real functions, so the notation for the complex conjugate has been omitted. It can be intuited from Equation 21 that $\alpha_{\mathrm{A}}^{1,2}$ is equivalent to the volume of the intersection of two spheres, as depicted in Figure 2.1.1.

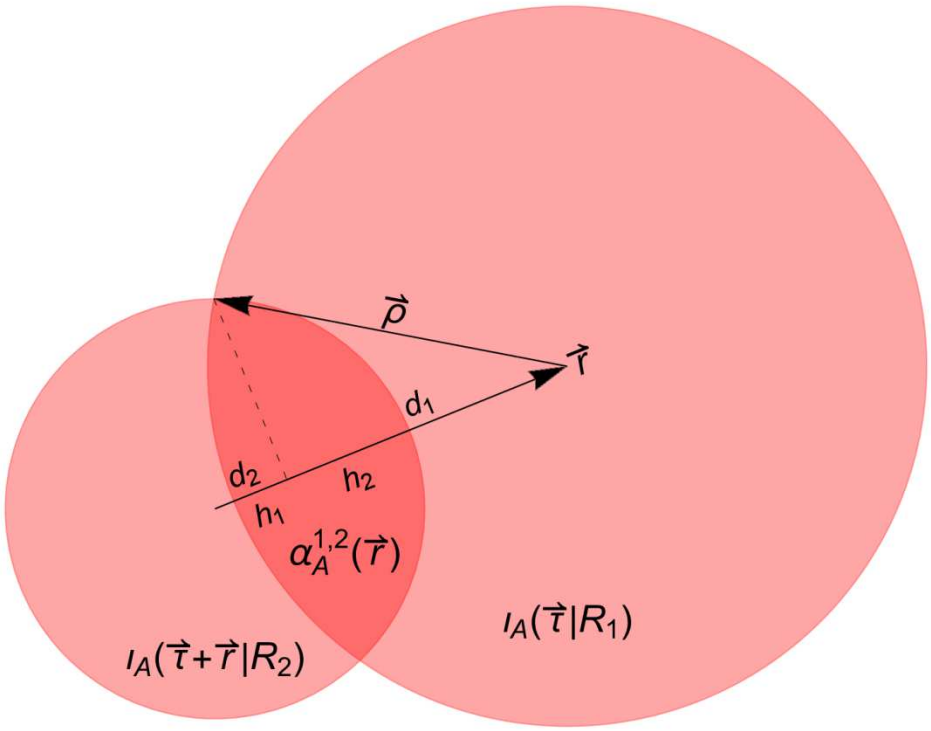

Figure 2.1.1 – Visualization of the functions $\iota_{\mathrm{A}}(\vec{\tau}|\boldsymbol{R_1})$ and $\iota_{\mathrm{A}}(\vec{\boldsymbol{r}}+\vec{\tau}|\boldsymbol{R_2})$ which depict uniformly dense spheres of radius $\boldsymbol{R_1}$ and $\boldsymbol{R_2}$ centered at the origin and $-\vec{\boldsymbol{r}}$, respectively. The volume corresponding to the intersection of these two spheres is equal to $\boldsymbol{\alpha_{\mathrm{A}}^{1,2}(\vec{r})}$.

The volume of the intersection can be calculated geometrically. A point $\vec{\rho} \in \mathbb{R}^3$ defined to be on the sphere surface in the plane of intersection must satisfy the following conditions:

$$\vec{\rho} \cdot \vec{\rho} = R_1^2 \tag{22}$$

$$(\vec{\rho} + \vec{r}) \cdot (\vec{\rho} + \vec{r}) = R_2^2 \tag{23}$$

Assuming there is an intersection of at least one point on the two spheres (i.e., $\|\vec{r}\| \leq R_1 + R_2$), the distance from the sphere centers to the plane of intersection can be found by substituting Equation 22 into 23 and expanding:

$$d_1 = -\vec{\rho} \cdot \vec{r}/\|\vec{r}\| = (\|\vec{r}\|^2 - R_2^2 + R_1^2)/2\|\vec{r}\| \tag{24}$$

$$d_2 = \|\vec{r}\| - d_1 = (\|\vec{r}\|^2 - R_1^2 + R_2^2)/2\|\vec{r}\| \tag{25}$$

The volume specified by $\alpha_{\mathrm{A}}^{1,2}$ can be expressed as the sum of two spherical caps:

$$\alpha_{\mathrm{A}}^{1,2}(\vec{r}) = \mathrm{V}_{\mathrm{cap}}(h_1|R_1) + \mathrm{V}_{\mathrm{cap}}(h_2|R_2) \tag{26}$$

where the volume of a spherical cap of height h taken from a sphere of radius R is:

$$\mathrm{V}_{\mathrm{cap}}(h|R) = \tfrac{\pi}{3}h^2(3R-h) \tag{27}$$

The heights of the two caps that comprise the intersection are:

$$h_1 = R_1 - d_1 = \frac{R_2^2 - (R_1 - \|\vec{r}\|)^2}{2\|\vec{r}\|}\mathrm{H}(R_1 + R_2 - \|\vec{r}\|) \tag{28}$$

$$h_2 = R_2 - d_2 = \frac{R_1^2 - (R_2 - \|\vec{r}\|)^2}{2\|\vec{r}\|}\mathrm{H}(R_1 + R_2 - \|\vec{r}\|) \tag{29}$$

Substituting Equations 27-29 into Equation 26 produces the following formula for $\alpha_{\mathrm{A}}^{1,2}$:

$$\alpha_{\mathrm{A}}^{1,2}(\vec{r}) = \frac{\pi(R_1 + R_2 - \|\vec{r}\|)\omega^{1,2}(\vec{r})}{12\|\vec{r}\|}\mathrm{H}(R_1 + R_2 - \|\vec{r}\|) \tag{30}$$

Where

$$\omega^{1,2}(\vec{r}) = \|\vec{r}\|^2 + 2\|\vec{r}\|(R_1 + R_2) - 3(R_1^2 + R_2^2) + 6R_1R_2 \tag{31}$$

In the case where $R_1 = R_2 = R$, Equation 30 can be substantially simplified:

$$\alpha_{\mathrm{A}}(\vec{r}) = \frac{\pi}{12}(4R + \|\vec{r}\|)(2R - \|\vec{r}\|)^2\mathrm{H}(2R - \|\vec{r}\|) \tag{32}$$

Using Equations 1, 16, 20, and 30, an exact expression can be found for $\alpha_F^{1,2}$:

$$\alpha_{\mathrm{F}}^{1,2}(\vec{r}) = \sum_{\vec{\tau}_1 \in \mathcal{S}_1} \sum_{\vec{\tau}_2 \in \mathcal{S}_2} \alpha_{\mathrm{A}}^{1,2}(\vec{r} - \vec{\tau}_2 + \vec{\tau}_1) \tag{33}$$

2.1.4 Accounting for Periodicity

Equation 33 is defined for systems of atomic coordinates in unrestricted infinite space. However, simulations are commonly performed using periodic boundary

conditions. Suppose the periodic box of a simulation is orthogonal with edges parallel to the $\hat{x}$, $\hat{y}$, and $\hat{z}$ directions, with the displacement from the corner at the origin to the furthest corner of the box defined by the vector $\vec{b} = b_x\hat{x} + b_y\hat{y} + b_z\hat{z} = [b_x, b_y, b_z]' \in \mathbb{R}^3$. The following function describes the displacement along a single periodic direction of length y:

$$\Delta(x|y) = \left[\left(x + \tfrac{1}{2}y\right)\right] \bmod y - \tfrac{1}{2}y \tag{34}$$

where mod represents the modulus after division (also known as the modulo operation).

For a given displacement vector $\vec{r} \in \mathbb{R}^3$, the smallest equivalent displacement vector in orthogonal periodic three-dimensional space defined by $\hat{x}$, $\hat{y}$, and $\hat{z}$ directions the would be:

$$\vec{\mathrm{D}}\left(\vec{r}\middle|\vec{b}\right) = \Delta(r_x|b_x)\hat{x} + \Delta\left(r_y\middle|b_y\right)\hat{y} + \Delta(r_z|b_z)\hat{z} \tag{35}$$

In the periodic case, the 2-point statistics can be normalized to yield the probability of finding a location within an atom of element 1 and another location within an atom of element 2 separated by $\vec{r}$:

$$\overline{\alpha}_{\mathrm{F}}^{1,2}\left(\vec{r}\middle|\vec{b}\right) = \frac{1}{b_x b_y b_z}\sum_{\vec{\tau}_1 \in \mathcal{S}_1}\sum_{\vec{\tau}_2 \in \mathcal{S}_2} \alpha_{\mathrm{A}}^{1,2}\left(\vec{\mathrm{D}}\left(\vec{r} - \vec{\tau}_2 + \vec{\tau}_1\middle|\vec{b}\right)\right) \tag{36}$$

2.1.5 Accounting for Non-Orthogonal Simulation Boxes

Equations 35 and 36 represent the case where the periodic simulation box is orthogonal. Figure 2.1.2 represents the case where the periodic directions are not orthogonal.

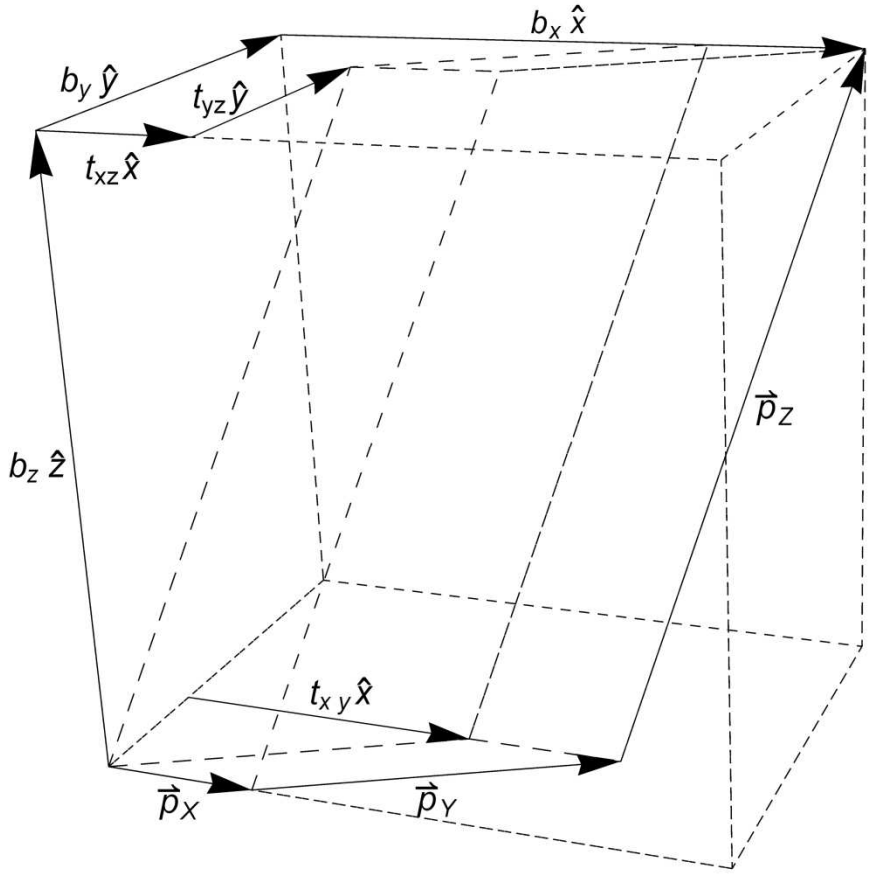

Figure 2.1.2 – Triclinic periodic unit cell with tilt factors t_{xy}, t_{xz}, and t_{yz}.,

In this case, $\vec{b}$ would represent the dimensions of the smallest orthogonal box containing the nonorthogonal simulation box and the box edges $\vec{p}_X$, $\vec{p}_Y$, and $\vec{p}_Z$ can be represented in terms of the tilt factors t_{xy}, t_{xz}, and as

$$\mathbf{P} = [\vec{p}_X \quad \vec{p}_Y \quad \vec{p}_Z\,] = \begin{bmatrix} b_x - |t_{xy}| - |t_{xz}| & t_{xy} & t_{xz} \\ 0 & b_y - |t_{yz}| & t_{yz} \\ 0 & 0 & b_z \end{bmatrix} \tag{37}$$

Here, the displacement vector in periodic space would be

$$\vec{\mathrm{D}}_{\dagger}(\vec{r}|\mathbf{P}) = \mathbf{P}\vec{\mathrm{D}}(\mathbf{P}^{-1}\vec{r}|\hat{x} + \hat{y} + \hat{z}) \tag{38}$$

and the 2-point statistics would be:

$$\overline{\alpha}^{1,2}_{\mathrm{F}\dagger}(\vec{r}|\mathbf{P}) = \frac{1}{|\det \mathbf{P}|}\sum_{\vec{\tau}_1 \in \mathcal{S}_1}\sum_{\vec{\tau}_2 \in \mathcal{S}_2} \alpha^{1,2}_{\mathrm{A}}\left(\vec{\mathrm{D}}_{\dagger}(\vec{r} - \vec{\tau}_2 + \vec{\tau}_1|\mathbf{P})\right) \tag{39}$$

where $\det \mathbf{P}$ is the determinant of $\mathbf{P}$.

2.1.6 Evaluation on a Discrete Grid

For the purposes of analysis, it is useful to calculate the 2-point statistics on a discretized grid. To do this, it is necessary to specify a vector $\vec{\ell} \in \mathbb{R}^3$ representing the dimensions of a voxel. If the 2-point statistics are to be represented by a 3rd-order tensor of size $\vec{G} = 2\vec{N} + 1$, where $\vec{N} = \left(N_x, N_y, N_z\right) \in \mathbb{Z}^3$ is a vector of integers representing the number of voxels from the center to the edge, then the set of allowable indices of that tensor would be

$$\Sigma_{\vec{N}} = \left\{(X, Y, Z) \in \mathbb{Z}^3 \middle| 0 \le X \le 2N_x, 0 \le Y \le 2N_y, 0 \le Z \le 2N_z\right\} \tag{40}$$

The 2-point statistics at a given set of indices $\vec{T} \in \Sigma_{\vec{N}}$ would then be

$$\mathrm{F}^{1,2}_{\Sigma_{\vec{N}}}\left[\vec{T}\right] = \overline{\alpha}^{1,2}_{\mathrm{F}}\left(\vec{X} \circ \vec{\ell} \middle| \vec{b}\right) \tag{41}$$

In Equation 41, the $\circ$ symbol represents element-wise multiplication. In addition to allowing for representation of the continuous function without the use of basis functions, discretization allows for the calculation of the 2-point statistics using the computationally efficient fast Fourier transform (FFT) due to the periodic nature of the data. The FFT and inverse FFT of a 3rd-order tensor with indices $\Sigma_{\vec{N}}$ are

$$\widetilde{\mathrm{M}}_{\Sigma_{\vec{N}}}\left[\vec{K}\right] = \sum_{\vec{T} \in \Sigma_{\vec{N}}} \mathrm{M}_{\Sigma_{\vec{\mathrm{N}}}}\left[\vec{T}\right] \exp\left(-2\pi i\left[\vec{K} \cdot \vec{T}\right] \oslash \vec{G}\right) \tag{42}$$

$$\mathrm{M}_{\Sigma_{\vec{N}}}\left[\vec{T}\right] = \frac{1}{G_x G_y G_z} \sum_{\vec{K} \in \Sigma_{\vec{N}}} \widetilde{\mathrm{M}}_{\Sigma_{\vec{\mathrm{N}}}}\left[\vec{K}\right] \exp\left(2\pi i\left[\vec{K} \cdot \vec{T}\right] \oslash \vec{G}\right) \tag{43}$$

The $\oslash$ symbol represents element-wise division and $\cdot$ represents the dot product. If one wishes to calculate the full discretized 2-pt statistics for atomic structures simulated with a periodic box of size $\vec{b}$, then the corresponding voxel size is simply:

$$\vec{\ell} = \vec{b} \oslash \vec{G} \tag{44}$$

The discretized sphere and atom center cross-correlations are, respectively:

$$\mathrm{A}_{\Sigma_{\vec{N}}}^{1,2}[\vec{T}] = \alpha_{\mathrm{A}}^{1,2}\left(\vec{\mathrm{D}}\left(\vec{T} \circ \vec{\ell}\middle|\vec{b}\right)\right) \tag{45}$$

$$\mathrm{C}_{\Sigma_{\vec{N}}}^{1,2}[\vec{T}] = \sum_{\vec{\tau}_1 \in \mathcal{S}_1} \sum_{\vec{\tau}_2 \in \mathcal{S}_2} \delta\left(\left\|\left[\vec{\mathrm{D}}\left((\vec{\tau}_2 - \vec{\tau}_1) - (\vec{T} - \vec{N}) \circ \vec{\ell}\middle|\vec{b}\right) \oslash \vec{\ell}\right]\right\|\right) \tag{46}$$

Here, $\lfloor z \rfloor$ is the floor function and the central peak is at $\vec{T} = \vec{N}$. To calculate the discretized statistics for the full image, first compute

$$\tilde{\mathrm{F}}_{\Sigma_{\vec{N}}}^{1,2}[\vec{K}] = \frac{1}{b_x b_y b_z} \tilde{\mathrm{C}}_{\Sigma_{\vec{N}}}^{1,2}[\vec{K}] \circ \tilde{\mathrm{A}}_{\Sigma_{\vec{N}}}^{1,2}[\vec{K}] \tag{47}$$

and then take the inverse FFT of the result.

2.1.7 Accounting for Changes in Simulation Box Size

For molecular dynamics simulations performed under the NPT ensemble (where number of atoms and the time-average of both pressure and temperature are fixed), the full statistics of each snapshot cannot be measured on grids with constant voxel size, since the fluctuating simulation box size cannot be perfectly divided into the same size voxels for every snapshot. To resolve this issue, partial sets of statistics are calculated. The full statistics are calculated for all displacement vectors that lie within an orthogonal box with 2 of the corners at $-\frac{1}{2}\vec{b}$ and $\frac{1}{2}\vec{b}$ and edges parallel to the $\hat{x}$, $\hat{y}$, and $\hat{z}$ directions. The partial statistics, on the other hand, are computed for displacement vectors that lie within a smaller box with 2 corners at $-\vec{c}$ and $\vec{c}$, where $\vec{c}$ is a cutoff vector. Here, the voxel size would be

$$\vec{\ell} = 2\vec{c} \oslash \vec{G} \tag{48}$$

The discretized partial statistics are no longer periodic. To correct for potential errors from the FFT due to edge effects, a slightly larger tensor should be used as an intermediary step. Here, the range of allowable indices is defined by:

$$\vec{N}_{\dagger} = \vec{N} + \left\lfloor \frac{1}{2}(R_1 + R_2)\vec{G} \oslash \vec{c} \right\rfloor \tag{49}$$

The set of allowable indices $\Sigma_{\vec{N}_{\dagger}}$ can be found by substituting Equation 49 into the value of $\vec{N}$ for Equation 40. The statistics tensor $\mathrm{F}^{1,2}_{\Sigma_{\vec{N}_{\dagger}}}[\vec{T}]$ can be calculated by substituting this result and Equation 48 into Equations 45-47 and taking the inverse FFT. It should be noted that this tensor has errors in the entries near the edges due to the fact that an FFT operation was performed on a non-periodic tensor. However, one can discards the entries near the edges, retaining the tensor $\mathrm{F}^{1,2}_{\Sigma_{\vec{N}}}[\vec{T}]$ by keeping only the entries of $\mathrm{F}^{1,2}_{\Sigma_{\vec{N}_{\dagger}}}[\vec{T}]$ whose indices lie within $\Sigma_{\vec{N}}$. This matrix is free from errors resulting from edge effects.

2.2 Identification and Characterization of Grain Boundary Structure

2.2.1 Centro-Symmetry Parameter

Atomistic simulations of grain boundaries represent a class of data sets where data-driven PSP linkages may prove particularly useful. Since a given property of interest for a GB (such as GB energy) is expected to be primarily derived from the atoms near the interface, a systematic method for classifying atoms as either "bulk" or "GB" atoms is required. The methods explained here are based upon the centro-symmetry parameter[51], which is defined for each atom as the sum of the squared magnitudes of the resultant of pairs of nearly oppositely facing displacement vectors to the atom's nearest neighbors. If N_1^K is the number of nearest neighbors for each atom in a given perfect crystal structure (12 for FCC), $\mathcal{Z}_1^K$ is the set of integers from 1 to N_1^K, and $\vec{R}_{i,a}^{(k)}$ is the displacement vector from the a^{th} atom in simulation i to its k^{th} nearest neighbor in periodic space, then the centro-symmetry parameter may be rigorously defined as:

$$c_{i,a} = \min\left\{\sum_{\mathcal{Z}\in\mathcal{S}}\left\|\sum_{k\in\mathcal{Z}}\vec{R}_{i,a}^{(k)}\right\|^2 : \mathcal{S}\subset\binom{\mathcal{Z}_1^{\mathrm{K}}}{2}, |\mathcal{S}| = \tfrac{1}{2}N_1^{\mathrm{K}}, \bigcup_{\mathcal{Z}\in\mathcal{S}}\mathcal{Z} = \mathcal{Z}_1^{\mathrm{K}}\right\} \tag{50}$$

Here, $\binom{\mathcal{Z}_1^{\mathrm{K}}}{2}$ is the set of all possible unordered pairs of numbers from 1 to N_1^K, and $\mathcal{S}$ represents a set of 6 such pairs (where $|\mathcal{S}|$ is the size or number of pairs in set $\mathcal{S}$ and the indexing set $\mathcal{Z}$ represents one such pair) subject to the constraint that each index from 1 to N_1^K appears in one and only one of these pairs..Verifying this constraint is extremely computationally expensive. For the case of FCC, there are $\frac{12!}{(2!)^6} = 7{,}484{,}400$

possible values of $\mathcal{S}$ that would have to be investigated for each atom. Thankfully, this is not necessary for most atomic structures of practical use. The 6 smallest values of $\left\|\sum_{k\in\mathcal{Z}} \vec{R}_{i,a}^{(k)}\right\|^2$ almost always correspond to unordered pairs representing nearly oppositely facing displacement vectors, such that the magnitude of the resultant of only $\frac{12!}{(2!)(10!)} = 66$ displacement vector pairs need to be computed for each atom. As such, the centro-symmetry parameter may be reasonably computed as:

$$c_{i,a} \cong \min\left\{\sum_{\mathcal{Z}\in\mathcal{S}^*}\left\|\sum_{k\in\mathcal{Z}} \vec{R}_{i,a}^{(k)}\right\|^2 : \mathcal{S}^* \subset \binom{\mathcal{Z}_1^{\mathrm{K}}}{2}, |\mathcal{S}^*| = ½N_1^{\mathrm{K}}\right\} \tag{51}$$

If the displacement vectors $\vec{R}_{i,a}^{(k)}$ are numbered as $\vec{R}_1, \dots, \vec{R}_{N_1^{\mathrm{K}}}$ such that $\vec{R}_j$ and $\vec{R}_{j+N_1^{\mathrm{K}}/2}$ are oppositely-facing vectors, $c_{i,a}$ may be expressed in the more traditional, though not mathematically rigorous, representation:

$$c_{i,a} = \sum_{j=1}^{N_1^{\mathrm{K}}/2}\left\|\vec{R}_j + \vec{R}_{j+N_1^{\mathrm{K}}/2}\right\|^2 \tag{52}$$

For the FCC case, the vectors $\vec{R}_1, \dots, \vec{R}_{12}$ are visualized in Figure 2.2.1.

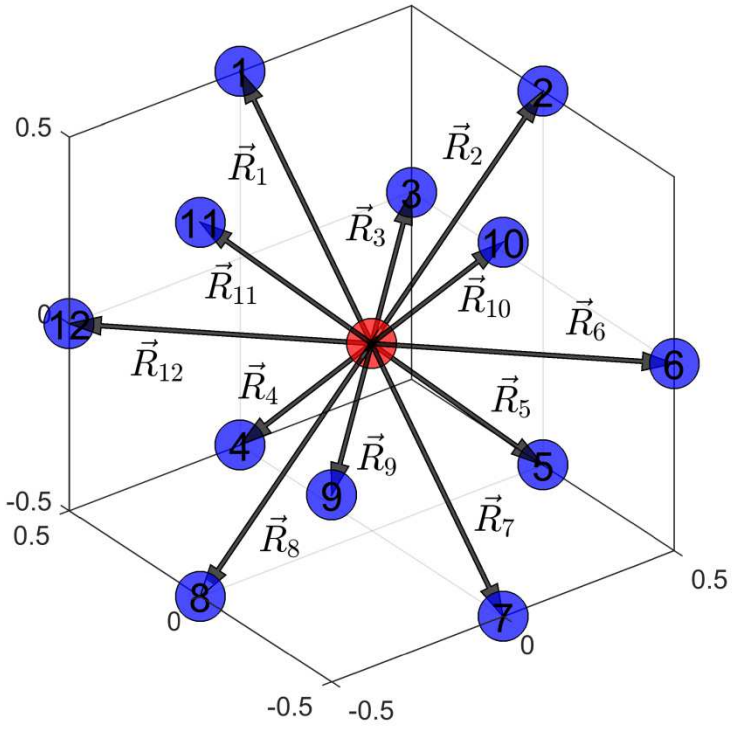

Figure 2.2.1 – Visualization of oppositely-facing displacement vectors for the centro-symmetry parameter of an atom in an FCC crystal.

2.2.2 Grain Boundary Atom Identification

The method used for classifying grain boundary atoms is based upon local regression[52]. Local regression is an application of weighted polynomial regression. The vector of coefficients $\underline{\hat{\beta}} = \left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right]^{\mathrm{T}}$ for a weighted polynomial regression fit of order p with responses v_a, weights w_a,and predictors s_a can be found from:

$$\left(\mathbf{S}_p^{\mathrm{T}}\mathbf{W}\mathbf{S}_p\right)\underline{\hat{\beta}} = \mathbf{S}_p^{\mathrm{T}}\mathbf{W}\underline{V} \tag{53}$$

where $\underline{V}$ is a vector whose a^{th} term is v_a, $\mathbf{S}_p$ is a matrix whose a^{th} row is $[1, s_a, \dots, s_a^p]$ and $\mathbf{W}$ is a diagonal matrix whose a^{th} term of the main diagonal is w_a. It should be noted that for sufficiently large values of p, scaling and mean-centering the columns of $\mathbf{S}_p$ may be necessary to avoid machine precision errors.

For local regression, the weights w_a are defined by the value of some kernel function with bandwidth h at s_a. In the case of a Gaussian kernel, which is the kernel employed in the GB atom identification method, the weights are:

$$w_a = \kappa^{\mathrm{N}}(s_a|h) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{s_a^2}{2h^2}\right) \tag{54}$$

where a good choice for h is the lattice constant of the crystal $a_0 = 4.05\text{Å}$ for the case of FCC Al.

The GB atom identification employs local quadratic regression (order $p = 2$). The responses for this regression problem are:

$$v_a = \sqrt{c_{i,a}} \tag{55}$$

where the square root of the centro-symmetry parameter was chosen for scaling purposes.

Consider the case where there is at least one GB in a simulation performed under periodic boundary conditions, and all GBs are perpendicular to the $\hat{y}$ direction. This requires that the tilt factors $t_{xy} = t_{yz} = 0$ such that the edges of the simulation box i as defined in Equation 37 are represented in the matrix

$$\mathbf{P}_i = \begin{bmatrix} b_x - |t_{xz}| & 0 & t_{xz} \\ 0 & b_y & 0 \\ 0 & 0 & b_z \end{bmatrix} \tag{56}$$

For this case, if $\vec{r}_{i,a}$ represents the Cartesian coordinates of atom a in simulation i, then the predictors of the regression problem are continuous functions of position $\vec{r} = r_x\hat{x} + r_y\hat{y} + r_z\hat{z}$, defined as:

$$s_a = \mathrm{S}_{i,a}(r_y) = \Delta(\vec{r}_{i,a} \cdot \hat{y} - r_y | b_y) \tag{57}$$

where the function Δ is defined in Equation 34.

Solving $\underline{\hat{\beta}}$ from Equation 53 using the values from Equations 54, 55, and 57 shows that the terms of $\underline{\hat{\beta}}$ can be expressed in terms of a continuous function of r_y:

$$\underline{\hat{\beta}} = [\hat{\beta}_{i0}(r_y), \hat{\beta}_{i1}(r_y), \hat{\beta}_{i2}(r_y)]^{\mathrm{T}} \tag{58}$$

where $\hat{\beta}_{ip}(r_y)$ is the regression coefficient for simulation i corresponding to the power-p term in $\mathbf{S}_2$.

In the GB atom identification procedure outlined here, the set of positions along the $\hat{y}$ direction corresponding to the GB centers and GB/bulk interfaces are defined, respectively, as:

$$\mathcal{Y}_{0i} = \{r_y | r_y \text{ is a local maximum of } \hat{\beta}_{i0}(r_y)\} \tag{59}$$

$$\mathcal{Y}_{2i} = \{r_y | r_y \text{ is a local maximum of } 2\hat{\beta}_{i2}(r_y)\} \tag{60}$$

The set of interface locations defined in Equation 60 correspond to the locations of the local maxima of the second derivative of local regression modeling equation. If the set of atom indices a for a given simulation i is $\mathcal{Z}_i^{\mathrm{A}}$ and we define functions representing the displacement along the $\hat{y}$ to the next highest and lowest GB interface, respectively, as:

$$\upsilon_{\mathrm{i}}(r_y) = \min\{\tilde{r}_y > 0 | \exists r_y^{\dagger} \in \mathcal{Y}_{2i} \text{ s.t. } \tilde{r}_y = \Delta(r_y^{\dagger} - r_y | b_y)\} \tag{61}$$

$$\lambda_{\mathrm{i}}(r_y) = \max\{\tilde{r}_y \leq 0 | \exists r_y^{\dagger} \in \mathcal{Y}_{2i} \text{ s.t. } \tilde{r}_y = \Delta(r_y^{\dagger} - r_y | b_y)\} \tag{62}$$

then we can define the set of atom indices a that are included in the GB:

$$\mathcal{Z}_i^{\mathrm{G}} = \{a \in \mathcal{Z}_i^{\mathrm{A}} | \exists r_y^* \in \mathcal{Y}_{0i} \text{ s.t. } \lambda_{\mathrm{i}}(r_y^*) \leq \mathrm{S}_{i,a}(r_y^*) < \upsilon_{\mathrm{i}}(r_y^*)\} \tag{63}$$

2.2.3 The Pair Correlation Function

The structure of the GB may be represented as a pair correlation function (PCF), which may be defined as a function of a distance t:

$$\gamma(t) = \frac{\mathrm{n}(t)}{4\pi t^2 n_0} \tag{64}$$

where n_0 is the number density of atoms in the perfect crystal (6.02×10^{-2} Å^{-3} for FCC Al) and $\mathrm{n}(t)$ is the average linear number density of atoms a distance t away from a given atom. It is clear to see that $n(t)$ can be defined in terms of a probability distribution function (PDF):

$$\mathrm{n}(t) = N\psi(t) \tag{65}$$

where N is the number of neighboring atoms and $\psi(t)$ is the probability density of finding a neighboring atom a distance away from a given atom.

If all N_i^A atoms in a given simulation are included in the PCF calculation, then $N = N_i^A - 1$. If $N = N_q^K$ where N_q^K is defined as the number of atoms in the 1^{st} through q^{th} set of nearest neighbors for a given crystal structure, then the PDF may be approximated using kernel density estimation (KDE):

$$\psi_{i,q}(t) = \frac{1}{N_q^K \left|\mathcal{Z}_i^{\mathrm{G}}\right|} \sum_{a \in \mathcal{Z}_i^{\mathrm{G}}} \sum_{k=1}^{N_q^K} \kappa\left(t - \left\|\vec{R}_{i,a}^{(k)}\right\| \middle| h_\dagger\right) \tag{66}$$

where κ is a kernel function with bandwidth $h_\dagger$.

The number density function would simply be

$$\mathrm{n}(t) = N_q^K \psi_{i,q}(t) \tag{67}$$

and the PCF may therefore be expressed in terms of $\psi_{i,q}(t)$ as:

$$\gamma_{i,q}(t) = \frac{N_q^K}{4\pi t^2 n_0} \psi_{i,q}(t) \tag{68}$$

If each atom is considered to be a uniformly dense sphere of radius R, a good candidate for a kernel function would be

$$\kappa(t|R,d) = \frac{\mathrm{A}(t|R,d)}{\frac{4}{3}\pi R^3} \tag{69}$$

In Equation 69, $\mathrm{A}(t|R,d)$ is the portion of the surface area of a sphere of radius t that lies within the volume of a second intersecting sphere of radius R, where the two sphere centers are separated by a distance d. then for a sufficiently large value of t, $\mathrm{A}(t|R,d)$ is approximately equal to the area of a cross-section of the second sphere a distance $t-d$ away from it's center:

$$\mathrm{A}(t|R,d) \cong \pi[R^2-(t-d)^2] \tag{70}$$

Using this value of $\mathrm{A}(t|R,d)$ and defining $u=t-d$ and $h_\dagger = R/\sqrt{5}$ produces the Epanechnikov kernel[53]:

$$\kappa(u|h_\dagger) = \kappa^{\mathrm{E}}(u|h_\dagger) = \begin{cases} \frac{3}{4h_\dagger\sqrt{5}}\left(1-\frac{u^2}{5h_\dagger^2}\right) & \text{for } (u/h)^2 < 5 \\ 0 & \text{otherwise} \end{cases} \tag{71}$$

As such, the Epanechnikov kernel is highly favorable for calculating PCFs using KDE. Without the approximation made in Equation 70, the kernel would change as a function of d:

$$\kappa(u|h_\dagger,d) = \left(1+\frac{u}{d}\right)\kappa^{\mathrm{E}}(u|h_\dagger) \tag{72}$$

The kernel representation of the PCF can be thought of as a generalization of a traditional binned PCF. For bins of width ℓ, the distance corresponding to the center of the j^{th} bin would be

$$t = \left(\tfrac{1}{2}+j\right)\ell \tag{73}$$

and the kernel function in Equation 66 would be $\kappa^{B}\left(u\middle|\ell/2\sqrt{3}\right)$ where κ^{B} is the box kernel function defined as:

$$\kappa^{B}(u|h') = \begin{cases} \frac{1}{2h'\sqrt{3}} & \text{for } |u/h'| < \sqrt{3} \\ 0 & \text{otherwise} \end{cases} \tag{74}$$

The PCF as defined in Equation 64 is a one-dimensional function measuring of distances to single neighboring atom. The PCF can be generalized to k dimensions representing distances to k neighboring atoms:

$$\gamma_{(k)}(t_1, \dots, t_k) = \frac{\mathrm{n}_{(k)}(t_1, \dots, t_k)}{n_0^k \prod_{i=1}^{k} 4\pi t_i^2} \tag{75}$$

where $\mathrm{n}_{(k)}$ is the average number density of atoms with neighbors a distance $t_1, \dots, t_k$ away. $\mathrm{n}_{(k)}$ may be expressed in terms of a probability $\psi_{(k)}$ of finding an atom with these neighbor distances:

$$\mathrm{n}_{(k)}(t_1, \dots, t_k) = \frac{N!}{(N-k)!}\psi_{(k)}(t_1, \dots, t_k) \tag{76}$$

2.3 Dimensionality Reduction and Regression using PCA and ASCA

2.3.1 Brief Discussion of ANOVA

ASCA (ANOVA Single Component Analysis)[47, 48], also called ANOVA-PCA[54], is a technique that generalizes PCA[49] to identify directions of maximum variance corresponding to sources of variance that are known prior by borrowing from the mathematics of ANOVA (Analysis of Variance)[55, 56]. A full overview of the uses and implementation of ANOVA is beyond the scope of this text; only the relevant aspects as pertaining to additive data decomposition are described here. A motivating example of two-way ANOVA would be a hypothetical hardness study of steels with different carbon compositions, each manufactured using the same set of heat treatments. In this example, carbon composition would be considered the first *factor* (called factor A); the i^{th} distinct composition being tested would be the i^{th} *level* of factor A. Heat treatment would be the second factor (B), with the j^{th} distinct heat treatment employed would the the j^{th} level of factor B. For each distinct combination of carbon composition and heat treatment, the k^{th} hardness test measurement would correspond to the k^{th} *replicate.*

Suppose you have some measurement y_{ijk} for the k^{th} replicate of the factor combination (i, j) for factors A and B, respectively. It is trivial to see that:

$$y_{ijk} - \bar{y}_{...} = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.}) \quad (77)$$

Here, $\bar{y}_{...}$ is the average over all measurements, $\bar{y}_{ij.}$ is the average measurement of all replicates for factor combination (i, j), and $\bar{y}_{i..}$ and $\bar{y}_{.j.}$ are the averages over all

measurements of the i^{th} and j^{th} levels of factors A and B, respectively. In this notation, a '$\cdot$' subscript indicates an average over that particular index. Equation 77 may be expressed symbolically as:

$$y_{ijk} - \hat{\eta} = \hat{\alpha}_i + \hat{\beta}_j + \hat{\omega}_{ij} + r_{ijk} \tag{78}$$

In Equation 78, $\hat{\alpha}_i$ and $\hat{\beta}_j$ are the 1-factor *main effects*, $\hat{\omega}_{ij}$ is the 2-factor *interaction* term, and r_{ijk} represents the *residuals*.

If there are K replicates for each combination of (i, j), I levels of factor A, and J levels of factor B, then according to ANOVA, the following equation must also hold true:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{ijk} - \hat{\eta}\right)^2 = KJ\sum_{i=1}^{I}\hat{\alpha}_i^2 + KI\sum_{j=1}^{J}\hat{\beta}_j^2 + K\sum_{I=1}^{I}\sum_{J=1}^{J}\hat{\omega}_{ij}^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}r_{ijk}^2 \tag{79}$$

2.3.2 Applicability of ANOVA Additive Decomposition to Covariance Matrices

In the language of ANOVA, the summations in Equation 79 are referred to as the *sum of squares*. The veracity of this equation may be demonstrated using linear transformations and the Kronecker product[57]. If $\mathbf{A}$ is an $m \times n$ matrix, and $\mathbf{B}$ is a $p \times q$ matrix, then the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ is an $mp \times nq$ matrix defined as

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix} \tag{80}$$

The Kronecker product has the following useful properties:

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}) \tag{81}$$

$$\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C} \tag{82}$$

$$(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C} \tag{83}$$

$$(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) \tag{84}$$

$$(n\mathbf{A}) \otimes \mathbf{B} = \mathbf{A} \otimes (n\mathbf{B}) = n(\mathbf{A} \otimes \mathbf{B}) \tag{85}$$

$$\left(\mathbf{A}^{\mathrm{T}} \otimes \mathbf{B}\right) \mathrm{vec}(\mathbf{C}) = \mathrm{vec}(\mathbf{BCA}) \tag{86}$$

In Equation 86, the function $\mathrm{vec}(\mathbf{C})$ represents the vectorization of the matrix $\mathrm{vec}(\mathbf{C})$ and $\mathbf{A}^{\mathrm{T}}$ is the transpose of $\mathbf{A}$. Returning to the original two-way ANOVA problem, the measurements taken can be represented in a block matrix:

$$\mathbf{Y}_{\mathrm{IJK}} = \begin{bmatrix} \underline{R}_{11} & \cdots & \underline{R}_{J1} \\ \vdots & \ddots & \vdots \\ \underline{R}_{1I} & \cdots & \underline{R}_{IJ} \end{bmatrix} \tag{87}$$

where the vector $\underline{R}_{ij}$ represents the set of replicates for the combination of factor levels (i, j) for factors A and B:

$$\underline{R}_{ij} = \left[y_{ij1}, \ldots, y_{ijk}, \ldots y_{ijK}\right]^{\mathrm{T}} \tag{88}$$

Define a matrix $\overline{\mathbf{Y}}_{\cdot\mathrm{J}\cdot}$ as a matrix where every y_{ijk} element of $\mathbf{Y}_{\mathrm{IJK}}$ is replaced with the value $\bar{y}_{\cdot j \cdot}$ corresponding to mean over all values with the same factor B level. This is equivalent to averaging over all entries in the same column of $\mathbf{Y}_{\mathrm{IJK}}$:

$$\overline{\mathbf{Y}}_{\mathrm{J}\cdot} = \frac{1}{I\times K}\mathbf{J}_{I\times K}\mathbf{Y}_{\mathrm{IJK}} = \left(\frac{1}{I}\mathbf{J}_I \otimes \frac{1}{K}\mathbf{J}_K\right)\mathbf{Y}_{\mathrm{IJK}} \tag{89}$$

where $\mathbf{J}_N$ is an $N \times N$ matrix of ones. In an analogous manner, the matrices $\overline{\mathbf{Y}}_{\mathrm{IJ}\cdot}$, $\overline{\mathbf{Y}}_{\mathrm{I}\cdot\cdot}$, and $\overline{\mathbf{Y}}_{\cdots}$, can be defined to be the matrices where the y_{ijk} elements of $\mathbf{Y}_{\mathrm{IJK}}$ are replaced with the corresponding $\bar{y}_{ij\cdot}$, $\bar{y}_{i\cdot\cdot}$, and $\bar{y}_{\cdots}$ values, respectively:

$$\overline{\mathbf{Y}}_{\mathrm{IJ}\cdot} = \left(\mathbf{I}_I \otimes \frac{1}{K}\mathbf{J}_K\right)\mathbf{Y}_{\mathrm{IJK}} \tag{90}$$

$$\overline{\mathbf{Y}}_{\mathrm{I}\cdot\cdot} = \left(\mathbf{I}_I \otimes \frac{1}{K}\mathbf{J}_K\right)\mathbf{Y}_{\mathrm{IJK}}\left(\frac{1}{J}\mathbf{J}_J\right) \tag{91}$$

$$\overline{\mathbf{Y}}_{\cdots} = \left(\frac{1}{I}\mathbf{J}_I \otimes \frac{1}{K}\mathbf{J}_K\right)\mathbf{Y}_{\mathrm{IJK}}\left(\frac{1}{J}\mathbf{J}_J\right) \tag{92}$$

Here, $\mathbf{I}_N$ is an $N \times N$ identity matrix.

Using Equations 89-92 and the properties demonstrated in Equations 82 and 83, the terms on the right-hand side of Equation 78 can be defined as matrices:

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{I}} = \overline{\mathbf{Y}}_{\mathrm{I}\cdot\cdot} - \overline{\mathbf{Y}}_{\cdots} = \left[\left(\mathbf{I}_I - \frac{1}{I}\mathbf{J}_I\right) \otimes \frac{1}{K}\mathbf{J}_K\right]\mathbf{Y}_{\mathrm{IJK}}\left(\frac{1}{J}\mathbf{J}_J\right) \tag{93}$$

$$\widehat{\boldsymbol{\beta}}_{\mathrm{J}} = \overline{\mathbf{Y}}_{\mathrm{J}\cdot} - \overline{\mathbf{Y}}_{\cdots} = \left[\frac{1}{I}\mathbf{J}_I \otimes \frac{1}{K}\mathbf{J}_K\right]\mathbf{Y}_{\mathrm{IJK}}\left(\mathbf{I}_J - \frac{1}{J}\mathbf{J}_J\right) \tag{94}$$

$$\widehat{\boldsymbol{\omega}}_{\mathrm{IJ}} = \overline{\mathbf{Y}}_{\mathrm{IJ}\cdot} - \overline{\mathbf{Y}}_{\mathrm{I}\cdot\cdot} - \overline{\mathbf{Y}}_{\mathrm{J}\cdot} + \overline{\mathbf{Y}}_{\cdots} = \left[\left(\mathbf{I}_I - \frac{1}{I}\mathbf{J}_I\right) \otimes \frac{1}{K}\mathbf{J}_K\right]\mathbf{Y}_{\mathrm{IJK}}\left(\mathbf{I}_J - \frac{1}{J}\mathbf{J}_J\right) \tag{95}$$

$$\mathbf{r}_{\mathrm{IJK}} = \overline{\mathbf{Y}}_{\mathrm{IJ}\cdot} - \overline{\mathbf{Y}}_{\cdots} = \left[\mathbf{I}_I \otimes \left(\mathbf{I}_K - \frac{1}{K}\mathbf{J}_K\right)\right]\mathbf{Y}_{\mathrm{IJK}}\left(\mathbf{I}_J - \frac{1}{J}\mathbf{J}_J\right) \tag{96}$$

The property depicted in Equation 86 can be used to convert the vectorizations of Equations 93-96 into linear transformations of the vectorization of $\mathbf{Y}_{\text{IJK}}$. This vectorization of $\mathbf{Y}_{\text{IJK}}$ is

$$\underline{Y} = \text{vec}(\mathbf{Y}_{\text{IJK}}) = \left[y_{111}, \ldots y_{11k}, \ldots y_{11K}, \ldots y_{i1k}, \ldots y_{I1K}, \ldots y_{ijk}, \ldots y_{IJK}\right]^{\text{T}} \tag{97}$$

and the vectorizations of Equations 93-96 are:

$$\text{vec}(\widehat{\boldsymbol{\alpha}}_{\text{I}}) = \left[\tfrac{1}{J}\mathbf{J}_J \otimes \left(\mathbf{I}_I - \tfrac{1}{I}\mathbf{J}_I\right) \otimes \tfrac{1}{K}\mathbf{J}_K\right]\underline{Y} = \mathbf{M}_{\text{A}}\underline{Y} \tag{98}$$

$$\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{J}}) = \left[\left(\mathbf{I}_J - \tfrac{1}{J}\mathbf{J}_J\right) \otimes \tfrac{1}{I}\mathbf{J}_I \otimes \tfrac{1}{K}\mathbf{J}_K\right]\underline{Y} = \mathbf{M}_{\text{B}}\underline{Y} \tag{99}$$

$$\text{vec}(\widehat{\boldsymbol{\omega}}_{\text{IJ}}) = \left[\left(\mathbf{I}_J - \tfrac{1}{J}\mathbf{J}_J\right) \otimes \left(\mathbf{I}_I - \tfrac{1}{I}\mathbf{J}_I\right) \otimes \tfrac{1}{K}\mathbf{J}_K\right]\underline{Y} = \mathbf{M}_{\text{A}\times\text{B}}\underline{Y} \tag{100}$$

$$\text{vec}(\mathbf{r}_{\text{IJK}}) = \left[\mathbf{I}_J \otimes \mathbf{I}_I \otimes \left(\mathbf{I}_K - \tfrac{1}{K}\mathbf{J}_K\right)\right]\underline{Y} = \mathbf{M}_{\text{e}}\underline{Y} \tag{101}$$

The vectorization of the left-hand side of Equation 78 may be expressed similarly as a linear transformation of $\underline{Y}$:

$$\text{vec}(\mathbf{Y}_{\text{IJK}} - \overline{\mathbf{Y}}_{\ldots}) = \underline{Y} - \text{vec}(\overline{\mathbf{Y}}_{\ldots}) = \left(\mathbf{I}_{I\times J\times K} - \tfrac{1}{I\times J\times K}\mathbf{J}_{I\times J\times K}\right)\underline{Y} = \mathbf{M}_0\underline{Y} \tag{102}$$

where matrix $\mathbf{M}_0$ performs a mean-centering operation on $\underline{Y}$. It is fairly trivial to show that

$$\mathbf{M}_0 = \mathbf{M}_{\text{A}} + \mathbf{M}_{\text{B}} + \mathbf{M}_{\text{A}\times\text{B}} + \mathbf{M}_{\text{e}} \tag{103}$$

As such, Equation 77 can be expressed in terms of linear transformations of $\underline{Y}$:

$$\mathbf{M}_0\underline{Y} = \mathbf{M}_\mathrm{A}\underline{Y} + \mathbf{M}_\mathrm{B}\underline{Y} + \mathbf{M}_{\mathrm{A}\times\mathrm{B}}\underline{Y} + \mathbf{M}_\mathrm{e}\underline{Y} \tag{104}$$

Since all $\mathbf{M}$ matrices are expressed in terms of $\mathbf{I}_N$, $\frac{1}{N}\mathbf{J}_N$, $\mathbf{I}_N$-$\frac{1}{N}\mathbf{J}_N$, all of which are idempotent, or Kronecker products of these matrices, then due to the properties illustrated in Equations 81 and 84, all the $\mathbf{M}$ matrices are also idempotent. Additionally, they are all symmetric due to similar reasons. This means that

$$\mathbf{M}^\mathrm{T}\mathbf{M} = \mathbf{M} \text{ for } \mathbf{M}_0, \mathbf{M}_\mathrm{A}, \mathbf{M}_\mathrm{B}, \mathbf{M}_{\mathrm{A}\times\mathrm{B}}, \text{and } \mathbf{M}_\mathrm{e} \tag{105}$$

As such, the left-hand side of Equation 79 may be expressed as

$$\mathrm{vec}\left(\mathbf{Y}_\mathrm{IJK} - \overline{\mathbf{Y}}_{\cdots}\right)^\mathrm{T} \mathrm{vec}\left(\mathbf{Y}_\mathrm{IJK} - \overline{\mathbf{Y}}_{\cdots}\right) = \underline{Y}^\mathrm{T}\mathbf{M}_0^\mathrm{T}\mathbf{M}_0\underline{Y} = \underline{Y}^\mathrm{T}\mathbf{M}_0\underline{Y} \tag{106}$$

Furthermore, also using the properties illustrated in Equations 81 and 84, it can be shown that $\mathbf{M}_\mathrm{A}$, $\mathbf{M}_\mathrm{B}$, $\mathbf{M}_{\mathrm{A}\times\mathrm{B}}$, and $\mathbf{M}_\mathrm{e}$ are all mutually orthogonal. For example,

$$\mathbf{M}_\mathrm{A}^\mathrm{T}\mathbf{M}_\mathrm{B} = \tfrac{1}{J}\mathbf{J}_J\left(\mathbf{I}_J - \tfrac{1}{J}\mathbf{J}_J\right) \otimes \left(\mathbf{I}_I - \tfrac{1}{I}\mathbf{J}_I\right)\tfrac{1}{I}\mathbf{J}_I \otimes \tfrac{1}{K}\mathbf{J}_K\tfrac{1}{K}\mathbf{J}_K = \mathbf{0}_{I\times J\times K} \tag{107}$$

where $\mathbf{0}_N$ is an $N \times N$ matrix of zeros. As such, by substituting Equation 103 into Equation 106, the following statement holds true, where each term represents a summation from Equation 79:

$$\underline{Y}^\mathrm{T}\mathbf{M}_0\underline{Y} = \underline{Y}^\mathrm{T}\mathbf{M}_\mathrm{A}\underline{Y} + \underline{Y}^\mathrm{T}\mathbf{M}_\mathrm{B}\underline{Y} + \underline{Y}^\mathrm{T}\mathbf{M}_{\mathrm{A}\times\mathrm{B}}\underline{Y} + \underline{Y}^\mathrm{T}\mathbf{M}_\mathrm{e}\underline{Y} \tag{108}$$

This proves the validity of the additive decomposition of the sum of squares in two-way ANOVA.

It is important to note that since $\mathbf{M}$ matrices represent linear transformations, they may operate on matrices as well as vectors. Though $\underline{Y}$ is defined as a vector, it can be replaced with a matrix and the conclusions from Equations 104 and 108 still hold true.

As such, in a manner analogous to the definition of $\underline{Y}$ in Equation 97, define a matrix $\mathbf{X}$, where each row $\underline{X}_{ijk}^{\mathrm{T}}$ represents a vector of some arbitrary length:

$$\mathbf{X} = \left[\underline{X}_{111}, \dots \underline{X}_{11k}, \dots \underline{X}_{11K}, \dots \underline{X}_{i1k}, \dots \underline{X}_{I1K}, \dots \underline{X}_{ijk}, \dots \underline{X}_{IJK}\right]^{\mathrm{T}} \tag{109}$$

It can then be therefore said that:

$$\mathbf{M}_0\mathbf{X} = \mathbf{M}_{\mathrm{A}}\mathbf{X} + \mathbf{M}_{\mathrm{B}}\mathbf{X} + \mathbf{M}_{\mathrm{A}\times\mathrm{B}}\mathbf{X} + \mathbf{M}_{\mathrm{e}}\mathbf{X} \tag{110}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{M}_0\mathbf{X} = \mathbf{X}^{\mathrm{T}}\mathbf{M}_{\mathrm{A}}\mathbf{X} + \mathbf{X}^{\mathrm{T}}\mathbf{M}_{\mathrm{B}}\mathbf{X} + \mathbf{X}^{\mathrm{T}}\mathbf{M}_{\mathrm{A}\times\mathrm{B}}\mathbf{X} + \mathbf{X}^{\mathrm{T}}\mathbf{M}_{\mathrm{e}}\mathbf{X} \tag{111}$$

The term on the left-hand side of represents the covariance matrix of $\mathbf{X}$ without the $\frac{1}{n-1}$ normalization term. It represents the covariance of $\mathbf{X}$ from all sources of variance. $\mathbf{X}^{\mathrm{T}}\mathbf{M}_{\mathrm{A}}\mathbf{X}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{M}_{\mathrm{B}}\mathbf{X}$ represent the covariance resulting only from changes in the levels of factors *A* and *B*, respectively, with all other sources of variance averaged out. $\mathbf{X}^{\mathrm{T}}\mathbf{M}_{\mathrm{A}\times\mathrm{B}}\mathbf{X}$ equals the covariance resulting from the interaction between changes in the levels of factors *A* and *B*, and $\mathbf{X}^{\mathrm{T}}\mathbf{M}_{\mathrm{e}}\mathbf{X}$ depicts the covariance from all remaining sources of variance not explained by the interaction or main effects. Equation 111 shows that the method governing the additive decomposition of the sum of squares in two-way ANOVA can also be applied to additively decompose a covariance matrix.

2.3.3 Principal Component Analysis (PCA)

Traditional PCA, a common technique for dimensional reduction where the j^{th} eigenvector corresponds to the direction with the j^{th} largest variance, is closely related to the singular value decomposition (SVD) of the mean-centered matrix:

$$\mathbf{M}_0\mathbf{X} = \mathbf{U}_0\mathbf{L}_0\mathbf{A}_0^{\mathrm{T}} \tag{112}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{M}_0\mathbf{X} = \mathbf{A}_0\mathbf{L}_0^2\mathbf{A}_0^{\mathrm{T}} \tag{113}$$

where $\mathbf{U}_0$ and $\mathbf{A}_0$ are orthonormal and $\mathbf{L}_0$ is a diagonal matrix of descending singular values ℓ_j^0. Here, j is the index corresponding to the j^{th} largest singular value. The columns of $\mathbf{A}_0$, $\underline{A}_j^0$, represent the PC eigenvectors. The matrix of PC scores is:

$$\mathbf{Z}_0 = \mathbf{U}_0\mathbf{L}_0 \tag{114}$$

where the columns of $\mathbf{Z}_0$, $\underline{Z}_j^0$, are the PC scores corresponding to the eigenvectors $\underline{A}_j^0$. Truncating the PCA representation of $\mathbf{M}_0\mathbf{X}$ at the k^{th} score/eigenvector yields the best possible rank-*k* approximation of the full dataset.

2.3.4 Principal Component Regression

The orthogonality of the PCA scores can be exploited to simplify the computation of principal component regression for predicting some vector of properties $\underline{P}$. For the purposes of calculation, it is useful to define a scalar term:

$$f_j = \underline{P}^{\mathrm{T}}\underline{U}_j^0 \tag{115}$$

where $\underline{U}_j^0$ is the j^{th} column of $\mathbf{U}_0$.

The regression coefficient corresponding to the PC scores $\underline{Z}_j^0$ for predicting $\underline{P}$ is simply:

$$c_j = f_j/\ell_j^0 \tag{116}$$

The PC scores corresponding to the largest singular values are not necessarily the best set of scores for predicting $\underline{P}$. Given that the PC scores are orthogonal, the best m-component PC model is equal to the sum of the m best 1-component PC models. The set of indices included in the best m-component PC model is:

$$\mathcal{S}_m = \begin{cases} \emptyset & m = 0 \\ \mathcal{S}_{m-1} \cup \left\{ \underset{j \notin \mathcal{S}_{m-1}}{\operatorname{argmax}} f_j^2 \right\} & m > 0 \end{cases} \tag{117}$$

where $\emptyset$ is an empty set and $\cup$ represents a union of sets. If $\bar{P}$ is the mean value of $\underline{P}$, then the estimated value of $\underline{P}$ predicted by the best m-component PC model is:

$$\underline{\hat{P}}_m = \bar{P} + \sum_{j \in \mathcal{S}_m} c_j \underline{Z}_j^0 = \bar{P} + \sum_{j \in \mathcal{S}_m} f_j \underline{U}_j^0 \tag{118}$$

The R^2 value and mean-squared error (MSE) and of the m-component model are

$$R^2 = \frac{\sum_{j \in \mathcal{S}_m} f_j^2}{\underline{P}^{\mathrm{T}} \mathbf{M}_0 \underline{P}} \tag{119}$$

$$MSE = \frac{\underline{P}^{\mathrm{T}}\mathbf{M}_0\underline{P} - \sum_{j\in\mathcal{S}_m} f_j^2}{L - m - 1} \tag{120}$$

where $L = I \times J \times K$ is the number of rows in the matrix $\mathbf{X}$.

2.3.5 ANOVA Single-Component Analysis (ASCA)

Equations 110 and 111 can be used to extend PCA to locate the directions of maximum variance from predetermined sources. In ASCA, an alternative to traditional PCA as outlined in Equations 112-114, SVD is performed on each term on the right-hand side of Equation 110 to yield the following results:

$$\mathbf{M}_0\mathbf{X} = \mathbf{U}_{\mathrm{A,B}}\mathbf{L}_{\mathrm{A,B}}\mathbf{A}_{\mathrm{A,B}}^{\mathrm{T}} = \mathbf{Z}_{\mathrm{A,B}}\mathbf{A}_{\mathrm{A,B}}^{\mathrm{T}} \tag{121}$$

where $\mathbf{U}_{\mathrm{A,B}}$, $\mathbf{L}_{\mathrm{A,B}}$, $\mathbf{A}_{\mathrm{A,B}}$, and $\mathbf{Z}_{\mathrm{A,B}}$ are equal to the following block matrices:

$$\mathbf{U}_{\mathrm{A,B}} = [\mathbf{U}_{\mathrm{A}} \quad \mathbf{U}_{\mathrm{B}} \quad \mathbf{U}_{\mathrm{A\times B}} \quad \mathbf{U}_{\mathrm{e}}] \tag{122}$$

$$\mathbf{L}_{\mathrm{A,B}} = \begin{bmatrix} \mathbf{L}_{\mathrm{A}} & 0 & \cdots & 0 \\ 0 & \mathbf{L}_{\mathrm{B}} & \ddots & \vdots \\ \vdots & \ddots & \mathbf{L}_{\mathrm{A\times B}} & 0 \\ 0 & \cdots & 0 & \mathbf{L}_{\mathrm{e}} \end{bmatrix} \tag{123}$$

$$\mathbf{A}_{\mathrm{A,B}} = [\mathbf{A}_{\mathrm{A}} \quad \mathbf{A}_{\mathrm{B}} \quad \mathbf{A}_{\mathrm{A\times B}} \quad \mathbf{A}_{\mathrm{e}}] \tag{124}$$

$$\mathbf{Z}_{\mathrm{A,B}} = [\mathbf{Z}_{\mathrm{A}} \quad \mathbf{Z}_{\mathrm{B}} \quad \mathbf{Z}_{\mathrm{A\times B}} \quad \mathbf{Z}_{\mathrm{e}}] = \mathbf{U}_{\mathrm{A,B}}\mathbf{L}_{\mathrm{A,B}} \tag{125}$$

where $\mathbf{M}_{\mathrm{A}}\mathbf{X} = \mathbf{U}_{\mathrm{A}}\mathbf{L}_{\mathrm{A}}\mathbf{A}_{\mathrm{A}}^{\mathrm{T}} = \mathbf{Z}_{\mathrm{A}}\mathbf{A}_{\mathrm{A}}^{\mathrm{T}}$ and likewise for the matrices with subscripts B, A × B, and e.

The scores $\mathbf{Z}_{\mathrm{A}}$, $\mathbf{Z}_{\mathrm{B}}$, $\mathbf{Z}_{\mathrm{A}\times\mathrm{B}}$, and $\mathbf{Z}_{\mathrm{e}}$ are all mutually orthogonal because their corresponding $\mathbf{M}$ matrices are all mutually orthogonal, as illustrated in Equation 107:

$$\mathbf{M}_{\mathrm{A}}^{\mathrm{T}}\mathbf{M}_{\mathrm{B}} = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{M}_{\mathrm{A}}^{\mathrm{T}}\mathbf{M}_{\mathrm{B}}\mathbf{X} = \mathbf{0} \Rightarrow \mathbf{A}_{\mathrm{A}}^{\mathrm{T}}\mathbf{Z}_{\mathrm{A}}^{\mathrm{T}}\mathbf{Z}_{\mathrm{B}}\mathbf{A}_{\mathrm{B}} = \mathbf{0} \Rightarrow \mathbf{Z}_{\mathrm{A}}^{\mathrm{T}}\mathbf{Z}_{\mathrm{B}} = \mathbf{0} \tag{126}$$

Since $\mathbf{Z}_{\mathrm{A,B}}$ is an orthogonal matrix, the simplified mathematics of PC regression described in Equations 115-120 hold true for ASCA as well as PCA.

2.3.6 Functional PCA and ASCA

PCA and ASCA can be applied to functions as well as matrices [58]. If there exists some vector of functions $\underline{\mathrm{F}}(t) = \left[\mathrm{f}_1(t), \ldots, \mathrm{f}_q(t)\right]^{\mathrm{T}}$, then the PCA or ASCA representation of these functions can be expressed as

$$\mathbf{M}\underline{\mathrm{F}}(t) = \mathbf{Z}\underline{\mathrm{A}}(\mathrm{t}) \tag{127}$$

where $\mathbf{M}$ is a mean-centering matrix from PCA or ASCA defined in Equations 98-102, $\mathbf{Z}$ is a $q \times n$ matrix of scores, and $\underline{\mathrm{A}}(t) = [\mathrm{a}_1(t), \ldots, \mathrm{a}_n(t)]^{\mathrm{T}}$ is a vector of orthonormal eigenfunctions. This can be solved by expressing $\underline{\mathrm{F}}(t)$ and $\underline{\mathrm{A}}(t)$ in terms of $\underline{\Phi}(t) = [\phi_1(t), \ldots, \phi_m(t)]^{\mathrm{T}}$, which is a vector of (not necessarily orthonormal) basis functions, such as B-splines:

$$\underline{\mathrm{F}}(t) = \mathbf{X}\underline{\Phi}(t) \tag{128}$$

$$\underline{\mathrm{A}}(t) = \mathbf{C}^{\mathrm{T}}\underline{\Phi}(t) \tag{129}$$

If the $m \times n$ matrix of eigenfunction coefficients $\mathbf{C}$ is defined such that $\mathbf{C} = \left[\underline{C}_1, \ldots, \underline{C}_n\right]$ and $\mathbf{C}_{\{i\}}$ corresponds to the first i columns of $\mathbf{C}$, then the eigenfunction coefficient vectors can be expressed as the solutions to the following optimization problem:

$$\underline{C}_i = \underset{\underline{C} \text{ s.t.} \mathbf{C}_{\{i-1\}}^{\mathrm{T}} \mathbf{K}_{(0)} \underline{C} = \underline{0}}{\operatorname{argmax}} \frac{\underline{C}^{\mathrm{T}} \mathbf{K}_{(0)} \mathbf{X}^{\mathrm{T}} \mathbf{M} \mathbf{X} \mathbf{K}_{(0)} \underline{C}}{\underline{C}^{\mathrm{T}} \mathbf{K}_{(0)} \underline{C}} \tag{130}$$

where $\underline{0}$ is a vector of zeros and the symmetric $m \times m$ weighting matrix $\mathbf{K}_{(0)}$ is defined as:

$$\mathbf{K}_{(0)} = \int \underline{\Phi}(t) \underline{\Phi}(t)^{\mathrm{T}} \mathrm{d}t \tag{131}$$

If $\mathbf{R}$ is defined as the upper right-hand triangular matrix from the Cholesky decomposition[59] of $\mathbf{K}_{(0)}$, i.e.,

$$\mathbf{R}^{\mathrm{T}} \mathbf{R} = \mathbf{K}_{(0)} \tag{132}$$

then the matrix $\mathbf{C}$ can be found from the results of the following singular value decomposition:

$$\mathbf{M} \mathbf{X} \mathbf{R}^{\mathrm{T}} = \mathbf{U} \mathbf{L} \mathbf{A}^{\mathrm{T}} \tag{133}$$

From here, the eigenfunction coefficients are simply:

$$\mathbf{C} = \mathbf{R}^{-1}\mathbf{A} \tag{134}$$

and the scores are:

$$\mathbf{Z} = \mathbf{MXK}_{(0)}\mathbf{C} \tag{135}$$

which can be simplified to:

$$\mathbf{Z} = \mathbf{UL} \tag{136}$$

2.3.7 Smoothed PCA/ASCA for Continuous Functions

If a smoothing hyper-parameter is desired, it is possible to incorporate a roughness penalty into Equation 130, where the columns of the penalized eigenfunction coefficient matrix $\mathbf{C}_\lambda = \left[\underline{C}_1^\lambda, \ldots, \underline{C}_n^\lambda\right]$ are defined by:

$$\underline{C}_i^\lambda = \underset{\underline{C}\ \text{s.t.}\ \mathbf{C}_{\{i-1\},\lambda}^{\mathrm{T}}\mathbf{K}_{(0)}\underline{C}=\underline{0}}{\operatorname{argmax}} \frac{\underline{C}^{\mathrm{T}}\mathbf{K}_{(0)}\mathbf{X}^{\mathrm{T}}\mathbf{MXK}_{(0)}\underline{C}}{\underline{C}^{\mathrm{T}}\left(\mathbf{K}_{(0)} + \lambda\mathbf{K}_{(2)}\right)\underline{C}} \tag{137}$$

where λ is a scalar roughness penalty chosen by the user, $\mathbf{C}_{\{i\},\lambda}$ corresponds to the first i columns of $\mathbf{C}_\lambda$, and $\mathbf{K}_{(2)}$ is defined as:

$$\mathbf{K}_{(2)} = \int \underline{\Phi}''(t)\underline{\Phi}''(t)^{\mathrm{T}}\mathrm{d}t \tag{138}$$

Equation 137 may be solved by selecting a matrix $\mathbf{B}_{i,\lambda}$ that has the following properties:

$$\mathbf{B}_{i,\lambda}^{\mathrm{T}}\mathbf{K}_{(0)}\mathbf{B}_{i,\lambda} = \mathbf{I}_{n-i} \tag{139}$$

$$\mathbf{B}_{i,\lambda}^{\mathrm{T}}\mathbf{K}_{(0)}\mathbf{C}_{\{i\},\lambda} = \mathbf{0} \tag{140}$$

The functional subspace defined by $\mathbf{B}_{i,\lambda}\underline{\Phi}(t)$ corresponds to the exclusion of the subspace defined by $\mathbf{C}_{\{i\},\lambda}\underline{\Phi}(t)$ from the subspace defined by $\mathbf{B}_{0,\lambda}\underline{\Phi}(t)$. To restrict solutions of Equation 137 to linear combinations of the unsmoothed eigenfunction coefficients, choose $\mathbf{B}_{0,\lambda} = \mathbf{C}$, where the columns of $\mathbf{C}$ correspond to solutions of Equation 130. To allow for additional smoothing, at the expense of the degree of dimensionality reduction, choose $\mathbf{B}_{0,\lambda} = \mathbf{R}^{-1}$, where $\mathbf{R}$ is defined in Equation 132. $\mathbf{B}_{i,\lambda}$ may be found from $\mathbf{B}_{0,\lambda}$ and $\mathbf{C}_{\{i\},\lambda}$ using a weighted Gram-Schmidt process where the inner products are weighted by $\mathbf{K}_{(0)}$.

Equation 137 may be solved in terms of the following singular value decomposition:

$$\mathbf{MXK}_{(0)}\mathbf{B}_{i-1,\lambda}\mathbf{R}_{i-1,\lambda}^{-1} = \mathbf{U}_{i,\lambda}\mathbf{L}_{i,\lambda}\mathbf{A}_{i,\lambda}^{\mathrm{T}} \tag{141}$$

where $\mathbf{R}_{i,\lambda}$ is defined by the following Cholesky decomposition:

$$\mathbf{R}_{i,\lambda}^{\mathrm{T}}\mathbf{R}_{i,\lambda} = \mathbf{B}_{i,\lambda}^{\mathrm{T}}\left(\mathbf{K}_{(0)} + \lambda\mathbf{K}_{(2)}\right)\mathbf{B}_{i,\lambda} \tag{142}$$

The unnormalized vector of coefficients from Equation 137 is

$$\underline{\breve{C}}_i^{\lambda} = \mathbf{B}_{i-1,\lambda}\mathbf{R}_{i-1,\lambda}^{-1}\underline{A}_1^{i,\lambda} \tag{143}$$

where $\underline{A}_1^{i,\lambda}$ is the first column of $\mathbf{A}_{i,\lambda}$ from Equation 141. The normalization can be found from:

$$\underline{C}_i^{\lambda} = \frac{1}{\sqrt{\breve{\underline{C}}_i^{\lambda \mathrm{T}} \mathbf{K}_{(0)} \breve{\underline{C}}_i^{\lambda}}} \breve{\underline{C}}_i^{\lambda} \tag{144}$$

The scores can be found by substituting $\mathbf{C}_{\lambda}$ into the value of $\mathbf{C}$ in Equation 135.

2.3.8 Smoothed PCA/ASCA for Uniformly Sampled Functions

The algorithm discussed in Section 2.3.7 can be adapted for the case where the data consists of functions sampled along fine, uniform intervals. In this case, if the data is represented by the $q \times m$ matrix $\mathbf{X} = \left[\underline{X}_1, \ldots, \underline{X}_m\right]$, then

$$\underline{X}_i = \underline{\mathrm{F}}(t_0 + (i-1)\ell) \tag{145}$$

where t_0 is the smallest point sampled and ℓ is the interval width.

The problem defined in Equation 137 can be solved by redefining $\mathbf{K}_{(2)}$ in terms of the second difference instead of the second derivative. If the first difference of a vector is the difference between consecutive elements in said vector, then the second difference is the difference between consecutive elements of the first difference. Now,

$$\mathbf{K}_{(2)} = \mathbf{D}_{(2)}^{\mathrm{T}} \mathbf{D}_{(2)} \tag{146}$$

where $\mathbf{D}_{(2)}$ is the $(m-2)\times m$ second difference transformation matrix. If $d_{ij}^{(2)}$ is the element in the ith row and jth column of $\mathbf{D}_{(2)}$, then

$$d_{ij}^{(2)} = \begin{cases} 1 & j = i \text{ or } j = i+2 \\ -2 & j = i+1 \\ 0 & \text{otherwise} \end{cases} \tag{147}$$

For the uniformly sampled case, the weighting matrix $\mathbf{K}_{(0)}$ is simply the identity matrix:

$$\mathbf{K}_{(0)} = \mathbf{I}_m \tag{148}$$

CHAPTER 3. INTERATOMIC POTENTIAL CLASSIFICATION FROM SIMULATED STRUCTURES

3.1 Overview

Here, the merits of a data-driven approach for addressing the challenge of mining and extracting core materials knowledge at the atomic scale are presented. The approach presented here is built on prior successes demonstrated for mesoscale representations of material internal structure, and involves three key steps: (i) discretization of the atomic structure, (ii) characterization of structure in the form of 2-point statistics, and (iii) representation of the structure in low-dimensional space using PCA. These novel protocols, applied on an ensemble of structure datasets output from MD simulations, have successfully classified the datasets based on several model input parameters such as the interatomic potential and the temperature used in the MD simulations.

3.2 Description of data

The data investigated here consisted of simulations of FCC aluminum performed at varying temperatures with various potentials housed at the NIST Interatomic Potentials Repository (http://www.ctcms.nist.gov/potentials). The interatomic potentials included in this study are summarized in Table 3.2.1, along with the appropriate references [60-79]. It is important to note that these calculations include some simulations well outside the intended usage of the interatomic potentials (e.g., using the pure elements of a potential only fit for use with compounds and thus they may not give the most accurate values for single-element atomic volumes). However, users often use interatomic potentials well

outside the range of where they were fit, and it is important to understand how that choice affects the answers obtained.

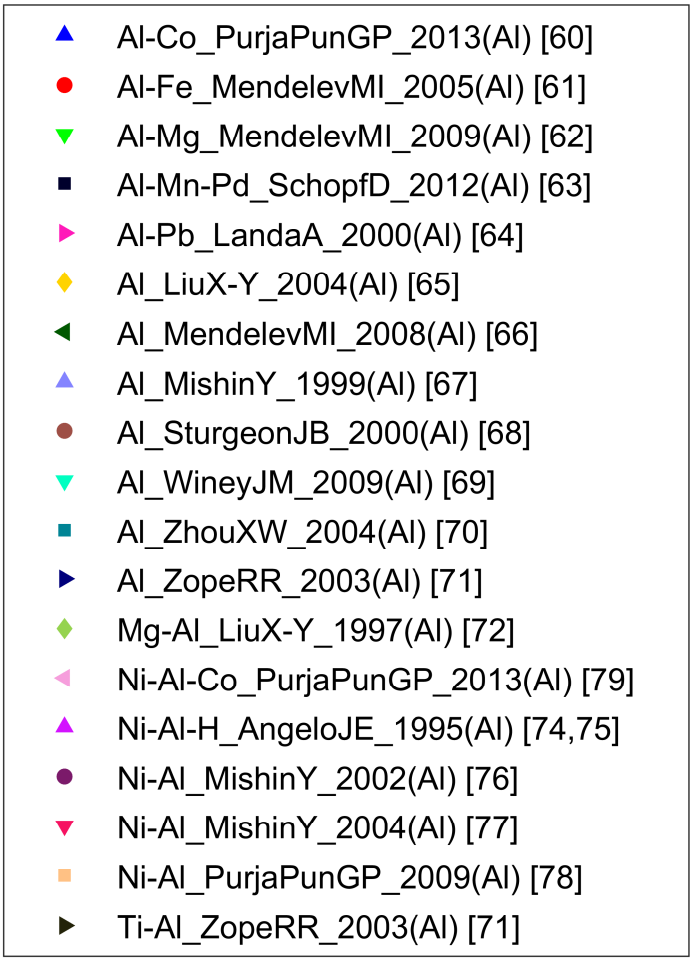

Table 3.2.1 – List of Al force fields used and their corresponding notation and references

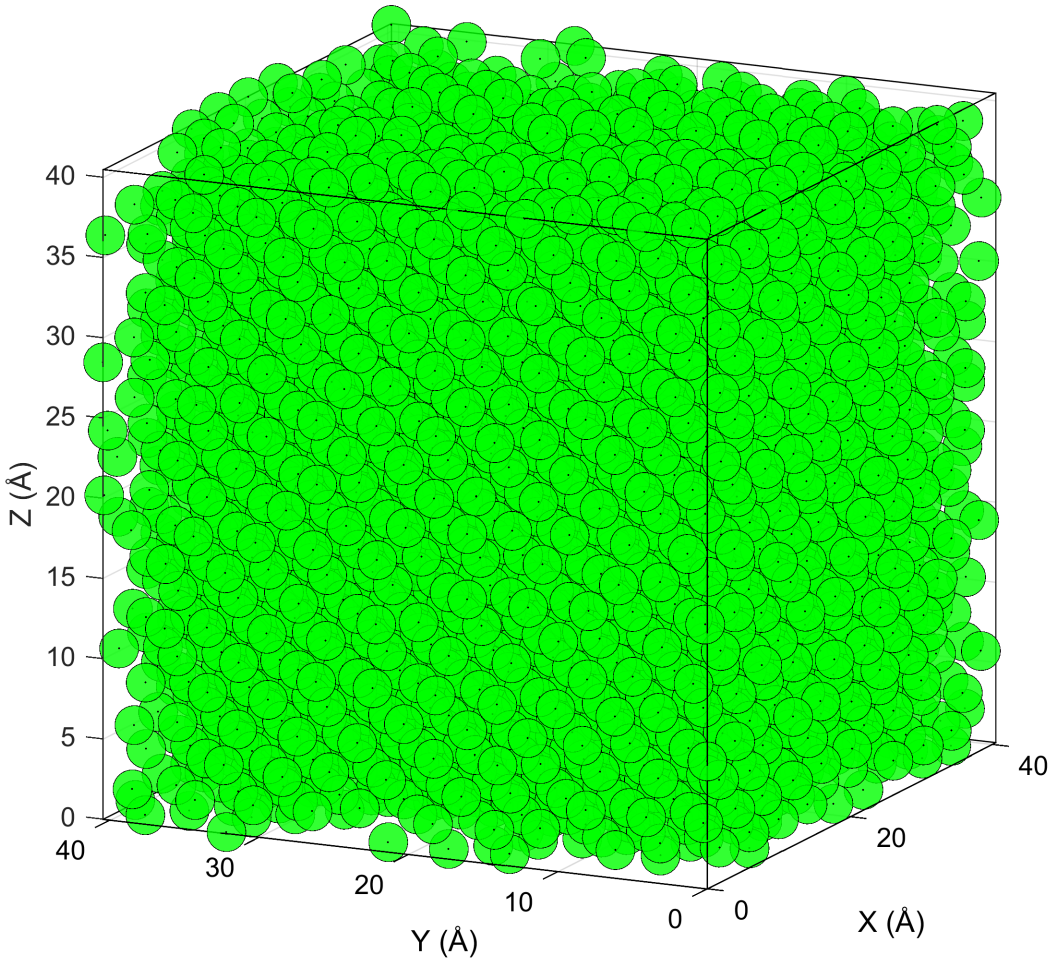

Figure 3.2.1 – Coordinates of a 4000 atom Al equilibrium simulation at 300 K at 10 ps using the force field "'Al-Pb_LandaA_2000." Dots represent atomic centers as generated by the simulation. For the purpose of 2-point statistics each atom was assigned a radius of 1.18 A, as depicted by the green circles. Though not clear in this figure, the structure is crystalline (face centered cubic) as expected.

After selection of the interatomic potential, the methodology for performing each simulation is as follows: (i) determine the 0 K equilibrium FCC lattice constant via a molecular statics simulation, (ii) create a 10 x 10 x 10 FCC unit cell (4000 atoms) using

the equilibrium lattice constant, (iii) create a uniform distribution of atomic velocities at the desired simulation temperature, and (iv) perform an isothermal-isobaric (NPT) simulation at the desired temperature for 2,000,000 time steps using a 1 fs time step. Data analysis described here takes place within the final 1,000,000 time steps. Instantaneous coordinates were recorded every 50,000 fs, and these were used in the analysis presented here. An example of one such snapshot is visualized in Figure 3.2.1.

3.3 Quantification of atomic structure

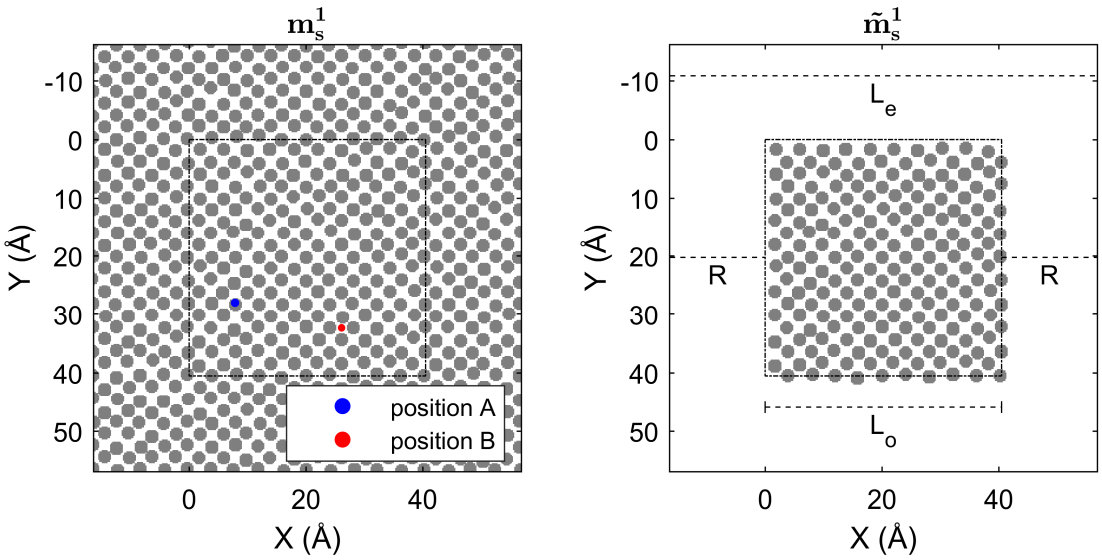

Figure 3.3.1 – Cross section corresponding to Z=20.24 Å of the corresponding discretized microstructure signals constructed in the novel protocols described in this paper. The full 3-D discretized images are used to calculate the 2-point statistics.

Structures were quantified in a manner similar to (though more computationally expensive than) the approach discussed in Section 2.1. (See Appendix A for full description). Briefly, the atomic coordinates of the atoms within the simulation box, as

well as the coordinates of atoms outside the simulation box in adjacent periodic boxes, were discretized in a grid with a predefined constant voxel size. All voxels whose center lied within an atomic radius (set as 1.18 Å) of an atom position were as assigned a value of 1; all over voxels were assigned a value of 0. The discretized 2-pt statistics were calculated using an FFT cross-correlation calculation (see Figure 3.3.1 for a visualization of the cross section of the discretized images included in the calculation).

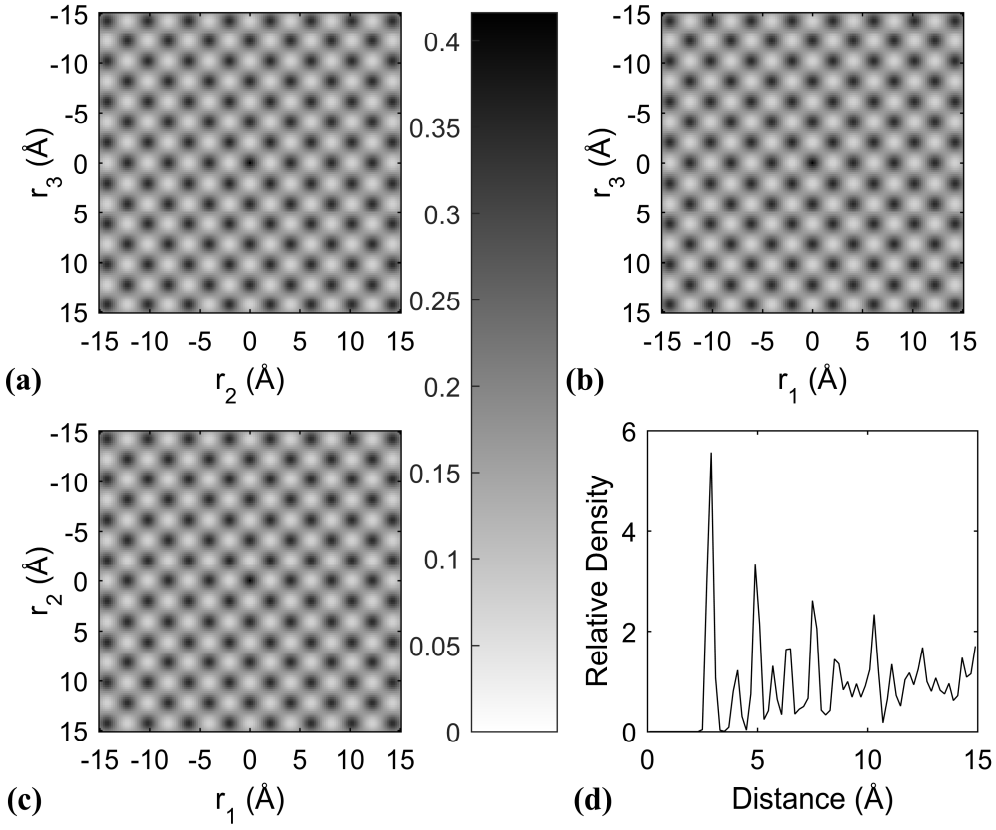

Figure 3.3.2 – The cross sections of the 2-point statistics of the data set shown in Figure 3.2.1 corresponding to (a) r_1=0, (b) r_2=0, and (c) r_3=0. The pair correlation function of this same structure is depicted in (d).

The only statistics retained corresponded to vectors that lie within a box defined by a cutoff vector $\vec{c}$ (see Section 2.1.7), with the rest discarded. The addition of extra atomic positions served two purposes: ensuring each snapshot (with varying simulation box size) can be discretized into an integer number of voxels, and to eliminate edge effects arising from the fact that the expanded box is no longer periodic.

Figure 3.3.2(a)-(c) illustrates two dimensional cross sections of the 2-pt statistics in $\mathbb{R}^3$. The pattern revealed in these corss sections is roughly consistent with the atomic positions corresponding to the FCC crystal structure, which is to be expected. Figure 3.3.2(d) presents the PCF, a more commonly used structure metric for MD simulations in the literature. The PCF corresponds to the spherically averaged 2-point statistics, and as such, depends strictly on the magnitude of the displacement vector, whereas the 2-point statistics retain both the magnitude and direction of this vector.

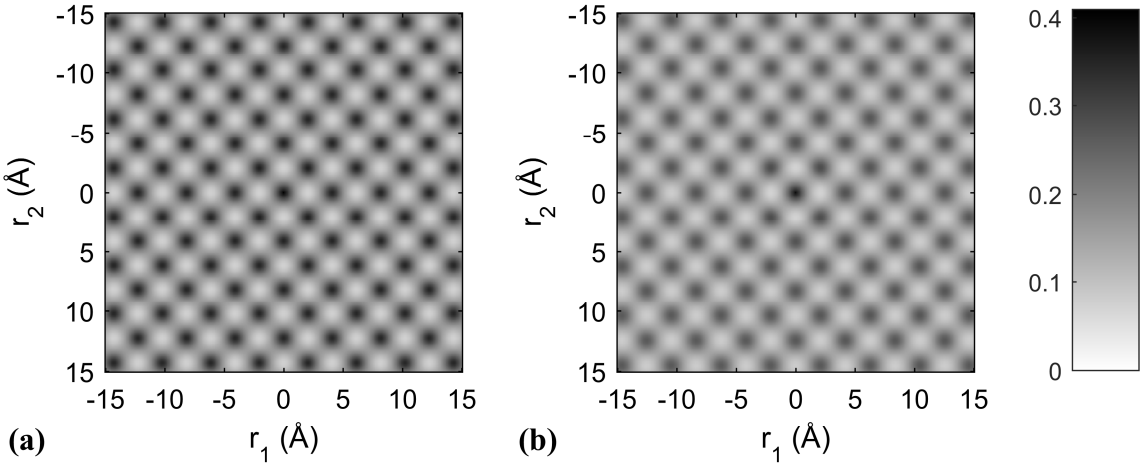

Figure 3.3.3 – r_3=0 cross sections of the 2-point statistics of the force field 'Al_SturgeonJB_2000(Al)' at (a) 300 K and (b) 900 K.

The greater intensity of the central peak in the 2-point statistics shown in Figure 3.3.2(a)-(c) in comparison to subsequent peaks is a product of the disorder due to thermal fluctuations (the greater the fluctuations, the greater the difference in intensity). For a perfectly periodic lattice, each peak will be of equal intensity. This effect is clearly noticeable in Figure 3.3.3, which corresponds to the 2-point statistics for the atomic structures at (a) 300 K and (b) 900 K. Also evident in these plots is the effect of thermal expansion; it can be seen from the peak positions of these two plots that the lattice constant at 900 K is greater than that of 300 K.

3.4 Low-Rank Model Construction

In this study, two different types of low rank models were investigated. First, simulations were grouped by equilibrium temperature, PCA (see Section 2.3.3) was performed separately for each group on the set of 2-point statistics characterizing the structures simulated by the 19 potentials at each of the 20 snapshots recorded, with particular emphasis on simulations performed at 300 K and 900 K. The distances in PC space corresponding to the first 3 principal components was used to classify the interatomic potentials into two distinct groups, along with a set of potentials whose behavior deviated greatly from the rest.

Secondly, PCA was performed using the time average of all snapshots from all simulations at all temperatures. In addition to further refining the groupings of the interatomic potentials, the PC scores from this study were analyzed in the in relation to the atomic volumes as a function of temperature.

3.5 Results and Discussion

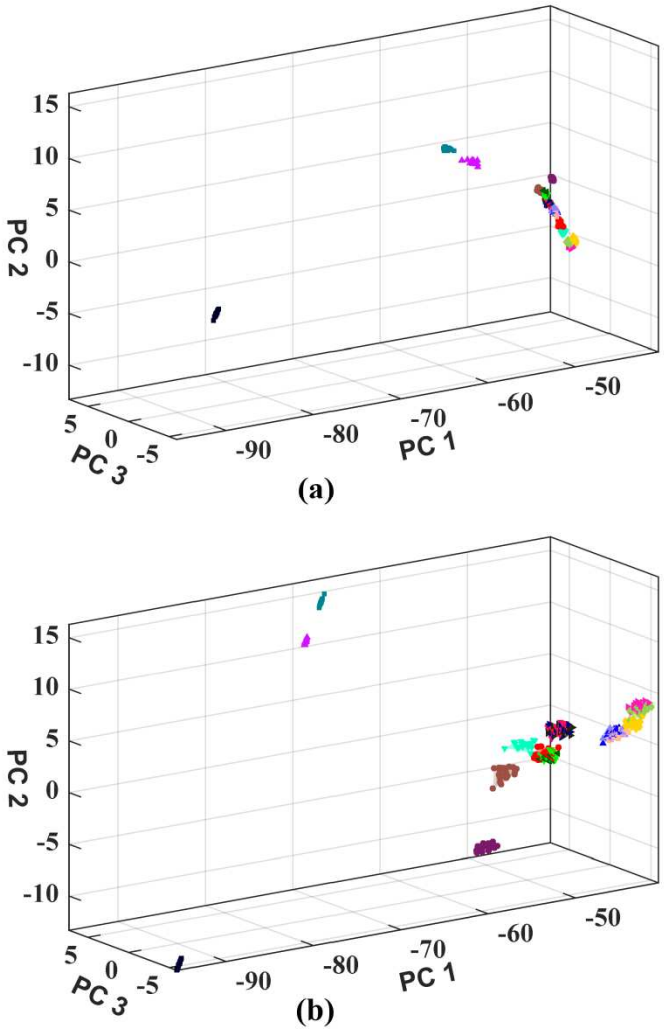

Figure 3.5.1 – The 2-point statistics every 50 ps from 1.05 ns to 2.0 ns of Al simulations using the force fields in Table 3.2.1 projected onto the first 3 principal components at 300 K (a) and 900 K (b) .

A visualization of the classification in PCA space of the MD datasets at 300K and 900K is depicted in Figure 3.5.1. Each data point in Figure 3.5.1 corresponds to the first three PC scores for each atomic structure included in the analyses. Despite a

dimensionality reduction of 119^3=1685159 to just three, this three PC score representation explains 99.8% of the total variance among the structures, echoing the substantial reduction in dimensionality experienced in analyses of mesoscale systems. As expected, the intra-class variance (reflected in the size of the cluster associated with each potential) is roughly equivalent for all potentials at the same temperature, and is substantially smaller than the inter-class variance.

The hierarchy of distances between clusters can be expressed as a dendrogram, which is depicted in Figure 3.5.2. Broadly, the PCA has identified the following clustering of potentials based on the differences in the structures produced by the MD simulations: the first group corresponds to the force fields referenced in [61, 64, 65, 69, 72], the second group corresponds to the force fields referenced in [60, 62, 66-68, 71, 73, 77, 78], including both force fields referenced in [71]. The four force fields referenced in [63, 70, 76] and [74, 75] are distinctly far away from the two groups identified above. The groupings of these results will be discussed in more detail in a later section.

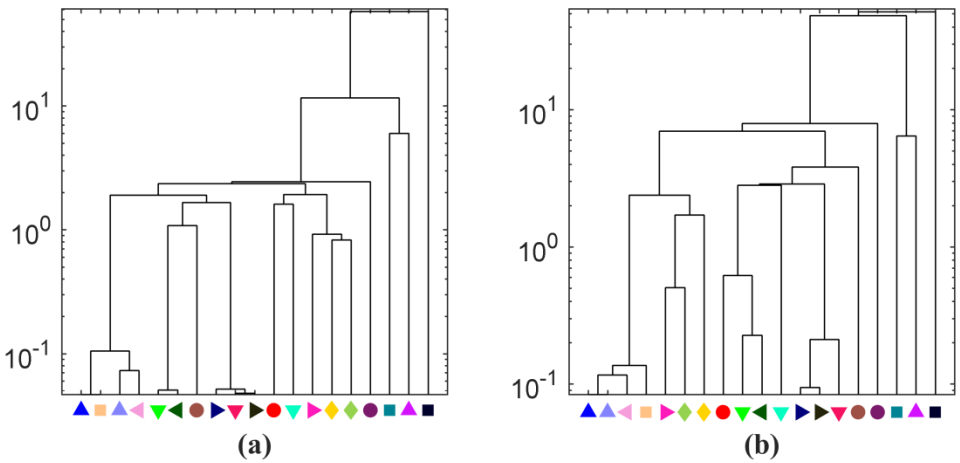

Figure 3.5.2 – The dendrograms of centroid distances of the data depicted in Figure 3 at 300 K (a) and 900 K (b).

Additional insights from the analysis presented here can be obtained from the plots of the PCs obtained in the analysis described above. Plots of the mean signal $\bar{f}_r$ and PCA eigenvectors φ_{ir} (for PC numbers i=1,2, and 3) are depicted in Figure 3.5.3. In these plots, black represents positive values and red represents negative. As such, a set of red and black spots in close proximity represents a s shift of the peak's position compared to the ensemble average. The overall plot of φ_{1r} therefore captures systematic shifts in the interatomic distances between any selected atom and its neighbors, with the shifts being higher for far away neighbors compared to those that are nearby. Therefore, φ_{1r} appears to capture well the overall volume differences among the snapshots of the atomic structure.

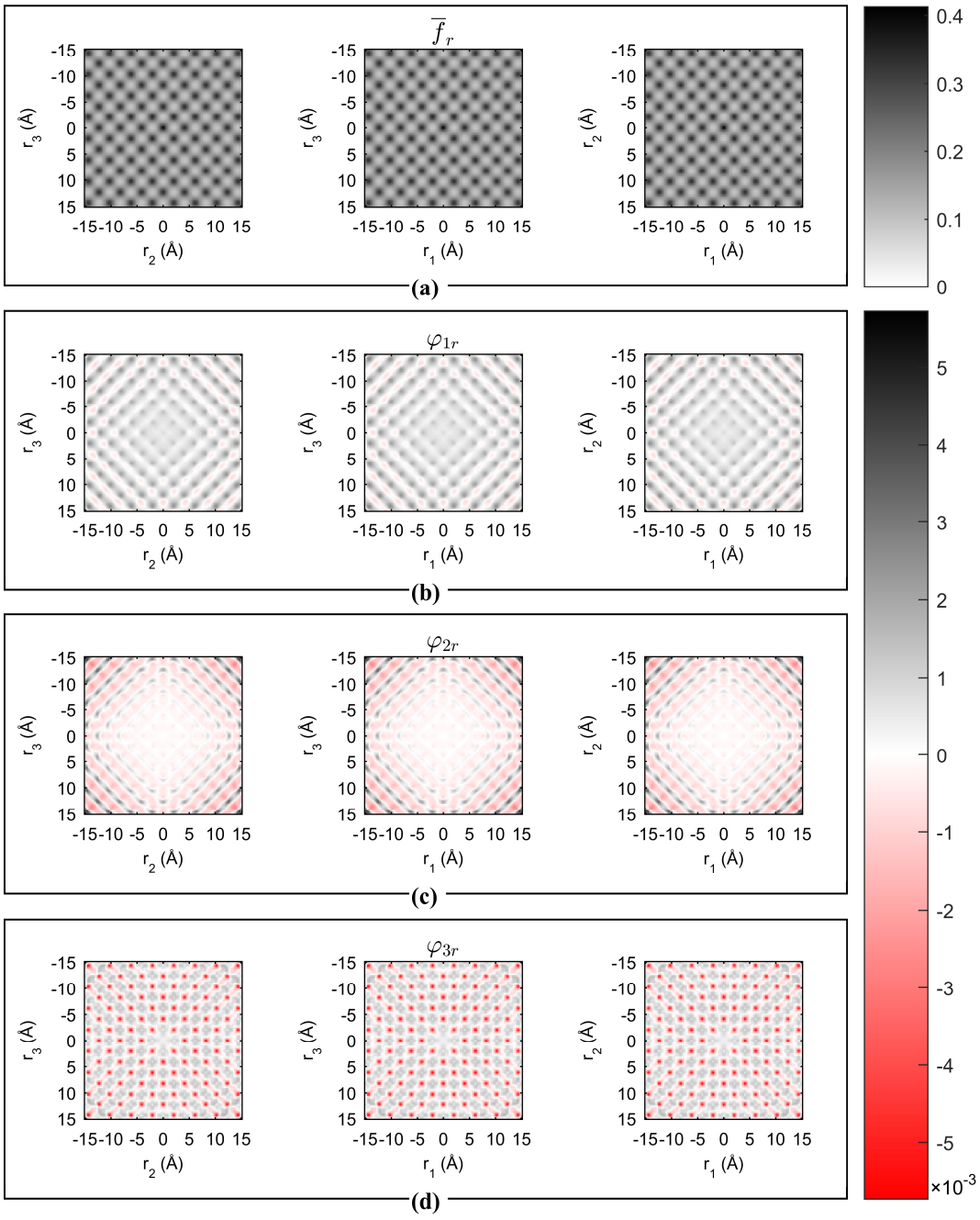

Figure 3.5.3 – Contour plots of the ensemble averaged spatial correlations and the PCA basis (eigenvectors) for the datasets shown in Figure 3.5.1(a), each shown as three orthogonal cross-sections.

For the analysis where PCA was performed on time-averaged 2-point statistics of all atomic structures at all temperatures, Figure 3.5.4 presents the PC scores as a function

of temperature. Of particular interest are the four force fields corresponding to References [63, 70, 76] and [74, 75] , which show significantly different behavior compared to the rest of the data sets. The force field used in Ref. [76] was strongly weighted to reproduce the properties of B2-NiAl, which may explain its poor behavior for pure aluminum. The other three interatomic potentials ([63, 70] and [74, 75]) were found to melt in the course of the simulations.

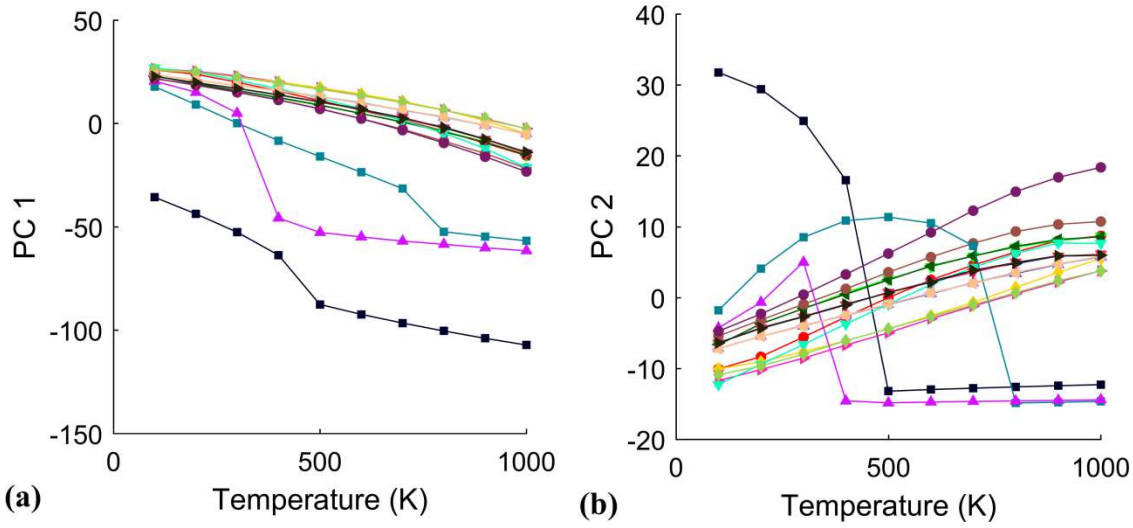

Figure 3.5.4 – The variation of (a) first principal component and (b) second principal component for the averaged 2-point statistics at each temperature. Only the mean 2-point statistics at each temperature for each force field were included in this PCA.

The difference in behavior for these four force fields is also evident in the temperature-dependent plots of average atomic volume, depicted in Figure 3.5.5. Furthermore, the groupings evident in this plot map directly to the groupings revealed in Figure 3.5.1(a). It should be stressed that PCA clustering is completely unsupervised. The fact that such analysis captures all of the significant differences in the predicted MD

structures supports the claim that the protocols used in this study produce high value, low dimensional, measures of the material structure.

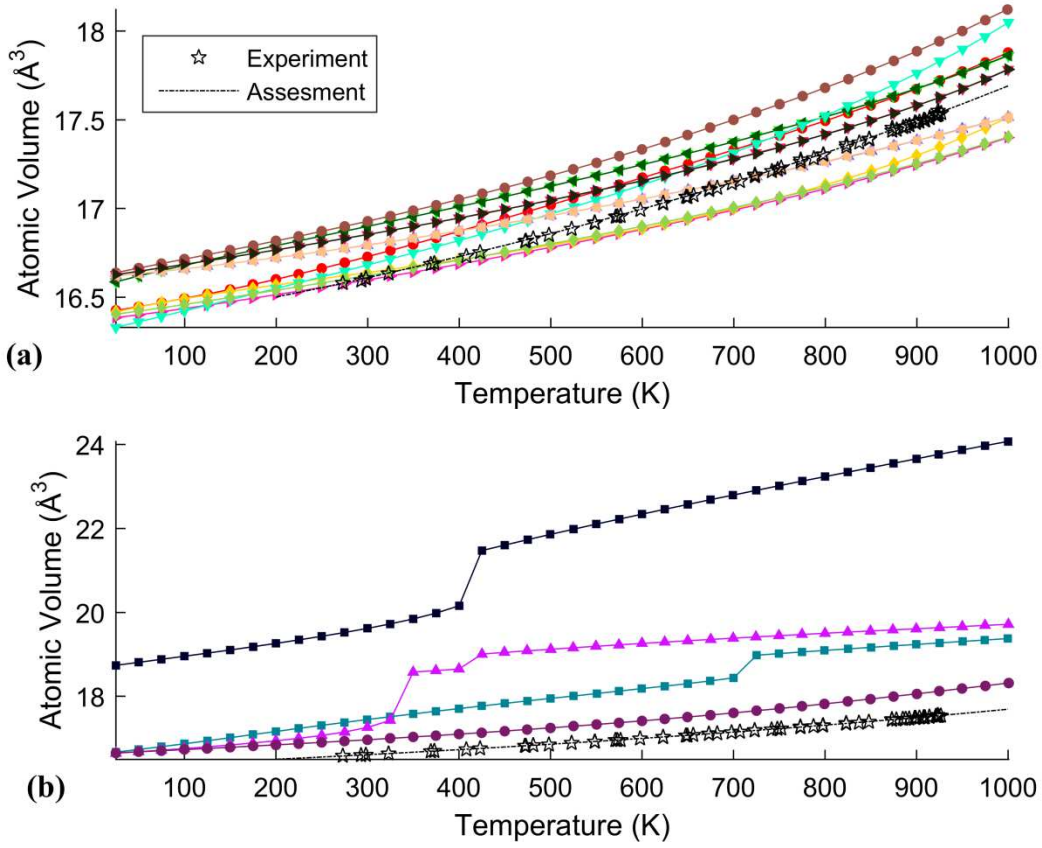

Figure 3.5.5 – Average atomic volumes from MD simulations of the (a) interatomic potentials closest to the experimental reference data, and (b) the four interatomic potentials exhibiting the largest deviation from the reference values. The discontinuities reflect phase changes associated with melting.

CHAPTER 4. EXTENSION OF PSP PARADIGM TO ATOMISTIC GB SIMULATIONS

4.1 Overview

In this study, it was shown that the "process-structure-property" (PSP) paradigm of materials science can be extended to atomistic grain boundary (GB) simulations through the development of a novel framework that addresses the objective identification of the atoms in the grain boundary regions using the centro-symmetry parameter and local regression, and the quantification of the resulting structure by a pair correlation function (PCF) derived from kernel density estimation (KDE). For asymmetric tilt GBs (ATGBs) in aluminum, models were successfully established connecting the GB macro degrees of freedom (treated as process parameters) and energy (treated as property) to a low-rank GB atomic structure approximation derived from principal component analysis (PCA) of the full ensemble of PCFs aggregated for this study. More specifically, it has been shown that the models produced in this study resulted in average prediction errors less than 13 mJ/m^2, which is less than the error associated with the underlying simulations when compared with experiments. This demonstration raises the potential for the development and application of PSP linkages from atomistic simulation datasets, and offers a powerful route for extracting high value actionable and transferrable knowledge from such computations.

The workflow developed and employed in this work for establishing the structure-property and process-structure linkages of interest is outlined in Figure 4.1.1. Broadly, this workflow depicts three main components: (i) low-rank quantification of the grain boundary atomic structure, (ii) extraction of a structure-property linkage, and (ii) extraction of process-structure linkages. These components are discussed below sequentially, explaining the rationale behind each step involved in each of these components.

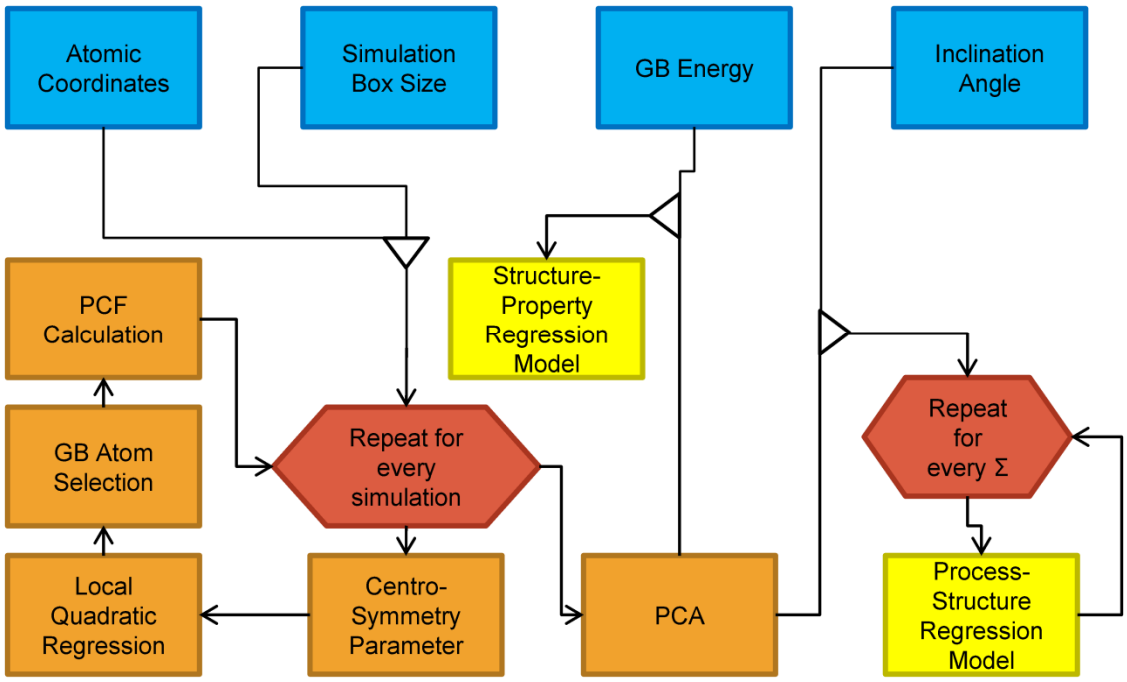

Figure 4.1.1 – Workflow employed in this study for establishing PSP linkages in simulated ATGBs.

4.2 Description of data

The dataset used in this study was produced by Tschopp et. al [40] and disseminated in an open repository hosted by the NIST Computational File

Repository[80]. The use of a publicly available dataset such as this allows multiple research groups (including ours) to apply different techniques and strategies, and to objectively compare the models produced. The reader is referred to previously published papers in literature[81] for details of how this dataset was generated.

The dataset available for the study contained a total of 106 molecular mechanics (MM) simulations that included ATGBs with Σ values of 3, 5, 9, 11, and 13 in aluminum (see Table 4.2.1), which reflect the level of coincidence of the atomic structure at the grain boundary. For example, a Σ of 3 indicates that 1/3 of the lattice sites of the two grain orientations that meet at the GB are coincident. All of the simulations employed periodic boundary conditions, with GB planes perpendicular to the $\hat{y}$ direction.[40]

	Misorientation Axis	No. of GBs	No. of atoms	Inclination Angle (θ)	GB Energy (mJ m^{-2})
Σ 3	[110]	26	284 – 8,096	0 – 90°	75 – 365
Σ 5	[001]	16	600 – 10,328	0 – 45°	465 – 542
Σ 9	[110]	27	852 – 14,624	0 – 90°	331 – 490
Σ 11	[110]	27	512 – 22,646	0 – 90°	151 – 431
Σ 13	[001]	10	1,040 – 7,668	0 – 45°	433 – 511

Table 4.2.1 – Details of grain boundary simulations used in this study[40, 80].

4.3 Identification of grain boundary atoms

A quantification of the structure of a GB must be independent of the volume of bulk crystal surrounding the GB, since there are more atoms in the bulk than at the GB yet these atoms contribute little to the GB energy. As such, a systematic way of classifying atoms as belonging to either the GB or the bulk must be employed. The method outline in Section 2.2 was employed for this study (see Figure 4.3.1a)

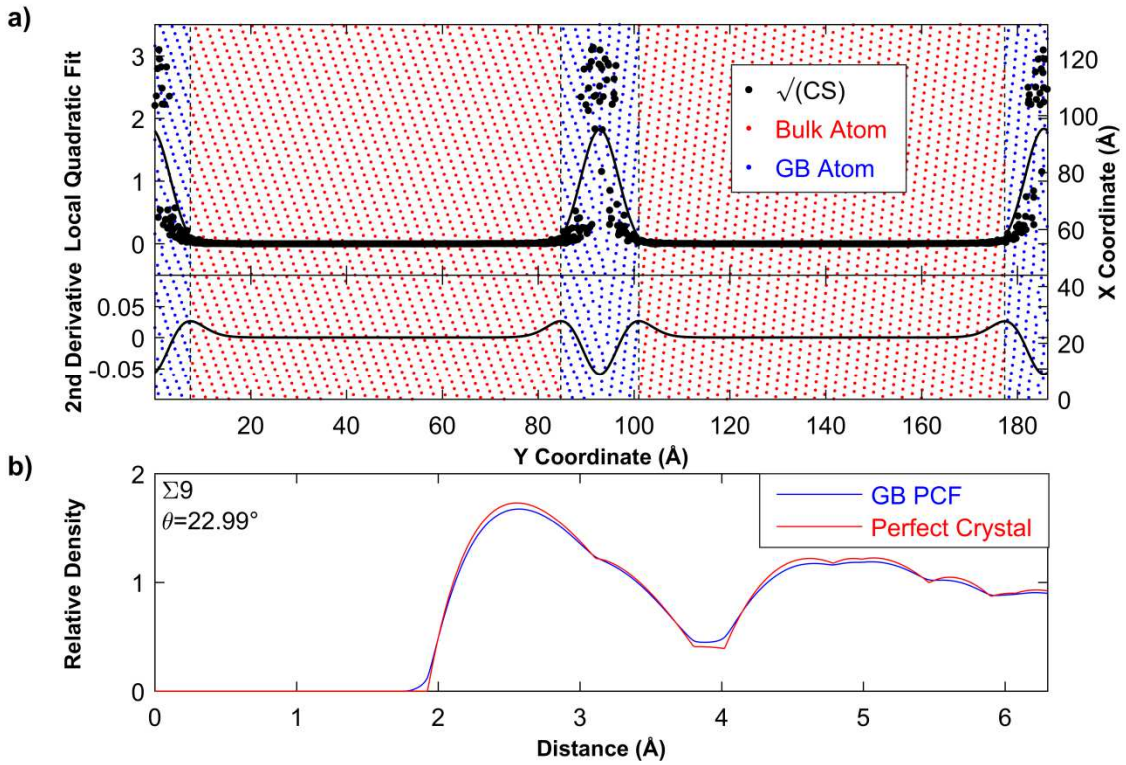

Figure 4.3.1 – Grain boundary selection procedure. (a) For a Σ9 asymmetric tilt grain boundary (ATGB) with an inclination angle (θ) of 22.99°, local quadratic regression fit (and corresponding local 2nd derivative) of the square root of the centrosymmetry parameter (CS) overlaid with atomic positions of grain boundary (GB) and bulk atoms. Dashed lines represent the interface between the GB and the bulk. (b) Pair correlation function (PCF) of this grain boundary in comparison to that of the perfect crystal.

4.4 Quantification of grain boundary structure

A rigorous characterization of atomic structure such as 2-point statistics would be informed by both the positions of the atoms and the orientation of the structure. The latter poses a challenge for GB simulations as each side of the GB corresponds to a different crystal orientation. Only the relative orientation of atoms with respect to other atoms would potentially be of interest, not the absolute orientation with respect to a reference frame, since any property predictive model based on the structure of a GB should be independent of the GB's orientation in space. Solutions to this problem may be found in

the field of computer vision[82], but for simulations such as this where energy is derived from the Embedded Atom Model (EAM), information retaining to orientation need not be retained in a structure characterization, allowing for structures to be quantified using PCFs, as explained in the next paragraph. KDE-derived PCFs with an Epanechnikov kernel, as described in Section 2.2.3 using neighbor distances up to the set of 7^{th} nearest neighbors ($q = 7, N_q^K = 134$) and sampled for 512 equally spaced points from 0 – 6.3Å, were used as the structure characterization metric in this study. Figure 4.3.1b depicts an example of one such PCF, in this case a Σ9 simulation with an inclination angle (θ) of 22.99°. For ATGBs, the inclination angle is the angle between the GB plane and the plane of reflection symmetry between the two crystal lattices.

Under EAM, the total energy for a group of atoms is:

$$E^{\mathrm{tot}} = \frac{1}{2}\sum_{a}\sum_{\tilde{a}\neq a}\Phi_{a\tilde{a}}(r_{a\tilde{a}}) + \sum_{a}\eta_a\left(\sum_{\tilde{a}\neq a}\rho_{\tilde{a}}(r_{a\tilde{a}})\right) \tag{149}$$

where E^{tot} is the energy of the system, $\Phi_{a\tilde{a}}$ is an interatomic pair potential, $\rho_{\tilde{a}}$ is the "atomic electron density" function, η_a is the embedding energy function, and $r_{a\tilde{a}}$ is the distance between atoms a and $\tilde{a}$. Inspection of the expression reveals that the interatomic distance, $r_{a\tilde{a}}$, is the fundamental variable of both summation terms. The PCF is a function of the probability distribution of $r_{a\tilde{a}}$, which provides a strong mathematical justification for its ability to accurately predict energies based on the Embedded Atom Model, and suggests that the accuracy of the approach will generalize to other similar potentials where the interatomic distance is the fundamental variable.[83, 84] However, it should be noted that a regression model constructed using PCFs calculated by a

traditional binning technique had relatively weak predictive power. This means that a predictive model must be robust against structural variance as pertaining to small local changes in atomic position, which are diminished by the smoothing parameter in KDE-derived PCFs.

4.5 Low-Rank Model Construction

The Epanechnikov kernel bandwidth $h_{\dagger}$ serves as a modeling hyperparameter, and the value chosen for the PCF calculation (0.42 Å) corresponds to the error minimum in the full PSP model and full dataset, as illustrated in Figure 4.5.1.

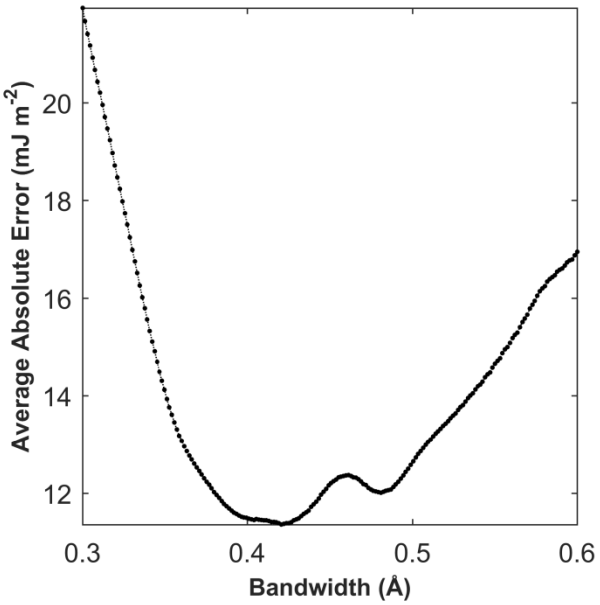

Figure 4.5.1 – Structure-property model error as a function of the PCF bandwidth.

After subtracting the mean PCF from the entire ensemble of discretely sampled PCFs of all simulations, PCA[49] was performed via the singular value decomposition

(see Section 2.3.3). The variance corresponding to the first 10 PCs are depicted in Figure 4.5.2a; Figure 4.5.2b depicts eigenvectors 1, 2, 3, and 6 of the dataset.

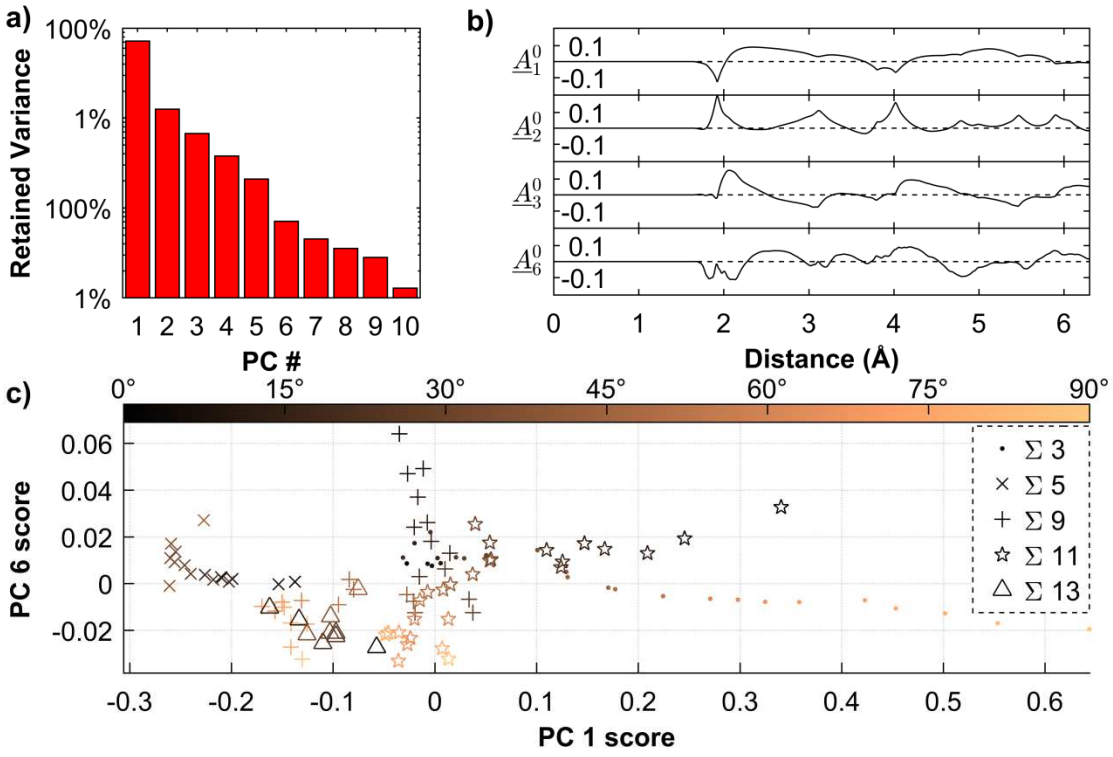

Figure 4.5.2 – Principal component analysis of GB PCFs. (a) Percentage of retained variance corresponding to the first 10 PCs on a logarithmic scale. (b) eigenvector ($\underline{A}_j^0$) associated with PC *j*, for *j* = 1, 2, 3, and 6, (c) Scores associated with PCs 1 and 6. θ is represented by the color scale.

In order to establish a quantitative structure-property relationship in this work, PC regression as outlined in was applied. For this data set, the predictors of best two-component structure-property regression model for predicting GB energy are the scores associated with PCs 1 and 6 (see Figure 4.5.2c).

Next, the model was analyzed to create a process-structure linkage. Employing the terms of a 3^{rd} order polynomial of θ as predictors, separate regression models were constructed for each Σ value to estimate the scores of PC 1 and 6.

4.6 Results and Discussion

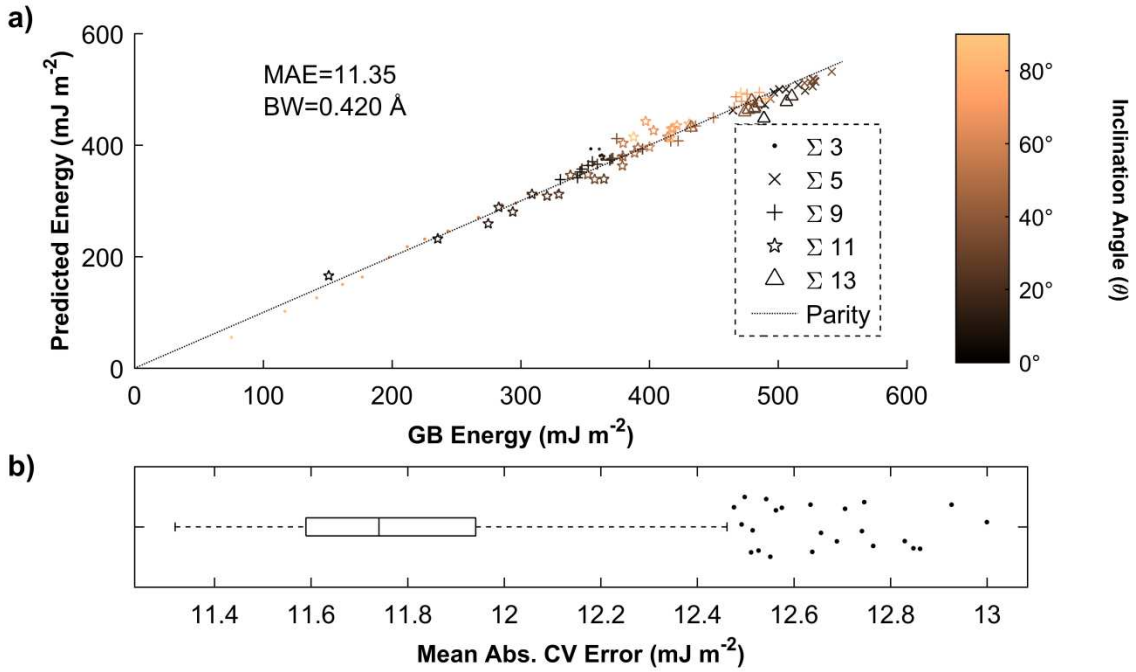

Figure 4.6.1 – Illustration of structure-property linkages. (a) Parity plot comparing the GB energies from atomistic simulations and the predicted values of GB energy from the 2-PC regression model. θ is represented by the color scale. (b) Box-Whisker plot of the mean absolute errors from 1000 instances of 3-fold cross-validation. The box represents the interquartile range, and the dashed 'whiskers' have a length 1.5 times that of the interquartile range; points outside this range represented as dots are considered outliers.

Figure 4.6.1a shows a parity plot of the GB energies predicted by the regression model plotted against the values computed from the full simulations. The regression coefficients for this model c_1, c_6 are -548.66 and -1070.89, respectively. The mean absolute value of the error in prediction is roughly 11.4 mJ/m^2. This is substantially less than the error in prediction of simulation versus experiment over a large range of θ as shown in the original dataset (see Ref. [40], Fig 1). Figure 4.6.1b shows a Box-Whisker plot of the mean absolute errors from 1000 instances of 3-fold cross-validation. This

shows that, even in the case of extreme outliers, the prediction error is still fairly small (< 13 mJ/m^2).

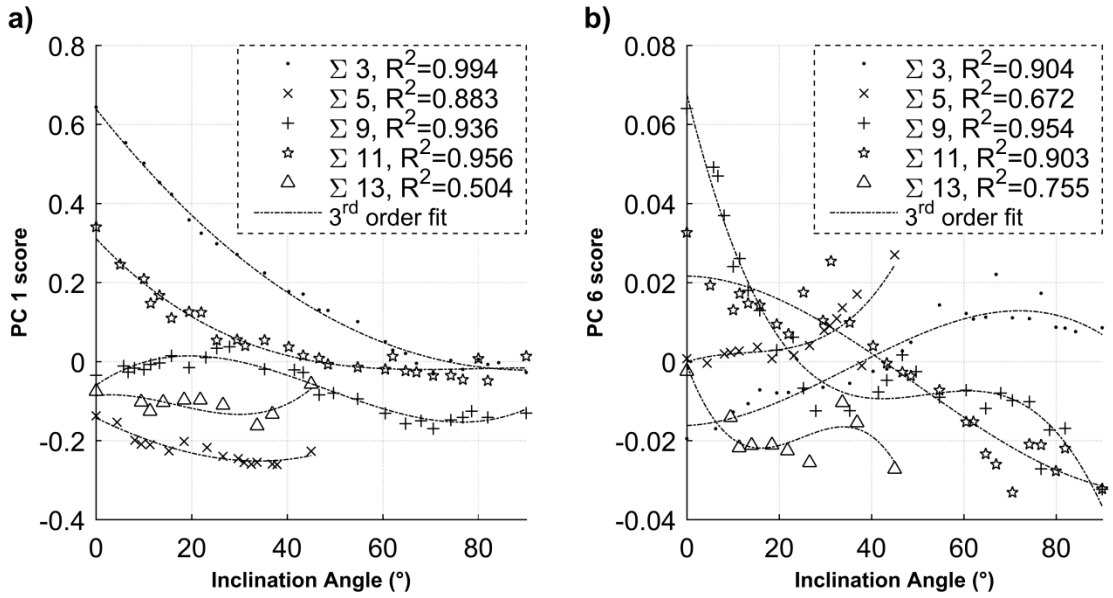

Figure 4.6.2 – Illustration of process-structure linkages. (a) the score as a function of θ and the model-predicted values for PC 1 and (b) PC 6. Points correspond to actual data and the 3rd order polynomial fit is indicated by the dashed line.

$PC\ score \approx A\theta^3 + B\theta^2 + C\theta + D$						
Coefficient	PC #	$\Sigma = 3$	5	9	11	13
A	1	8.71×10^{-8}	4.77×10^{-7}	1.83×10^{-6}	-6.82×10^{-7}	-5.16×10^{-6}
	6	1.45×10^{-7}	5.95×10^{-7}	-6.10×10^{-7}	1.09×10^{-7}	1.9×10^{-6}
B	1	7.29×10^{-5}	6.21×10^{-5}	-2.64×10^{-4}	1.64×10^{-4}	4.4×10^{-4}
	6	-2.38×10^{-5}	-2.31×10^{-5}	9.46×10^{-5}	-1.62×10^{-5}	-1.14×10^{-4}
C	1	7.7×10^{-5}	5.87×10^{-3}	8.25×10^{-3}	-1.29×10^{-2}	-9.75×10^{-3}
	6	7.17×10^{-4}	3.78×10^{-4}	-4.73×10^{-3}	-1.67×10^{-5}	1.84×10^{-3}
D	1	-2.16×10^{-2}	-1.43×10^{-1}	-5.98×10^{-2}	3.11×10^{-1}	-6.9×10^{-2}
	6	6.84×10^{-3}	-6.99×10^{-5}	6.76×10^{-2}	2.16×10^{-2}	-2.54×10^{-2}

Table 4.6.1 – Regression coefficients of the process-structure models.

Figure 4.6.2 shows the PC 1 and 6 scores along with the scores predicted from the regression models; the regression coefficients are listed in Table 4.6.1. For both PC 1 and 6, the regression fit for Σ =5 and 13 are much poorer than the fit for Σ = 3, 9, and 11.

ATGBs with for Σ = 5 and 13 and Σ = 3, 9, and 11 have misorientation axes of [001] and [110], respectively. The misorientation axis is the axis about which the lattices on either side of the GB are rotated to bring them into coincidence. This suggests that different misorientation axes influence the orientation of the PC vectors in a complex manner that is not fully described in this simple model. However, this added complexity does not manifest itself in the previously described structure-energy relation.

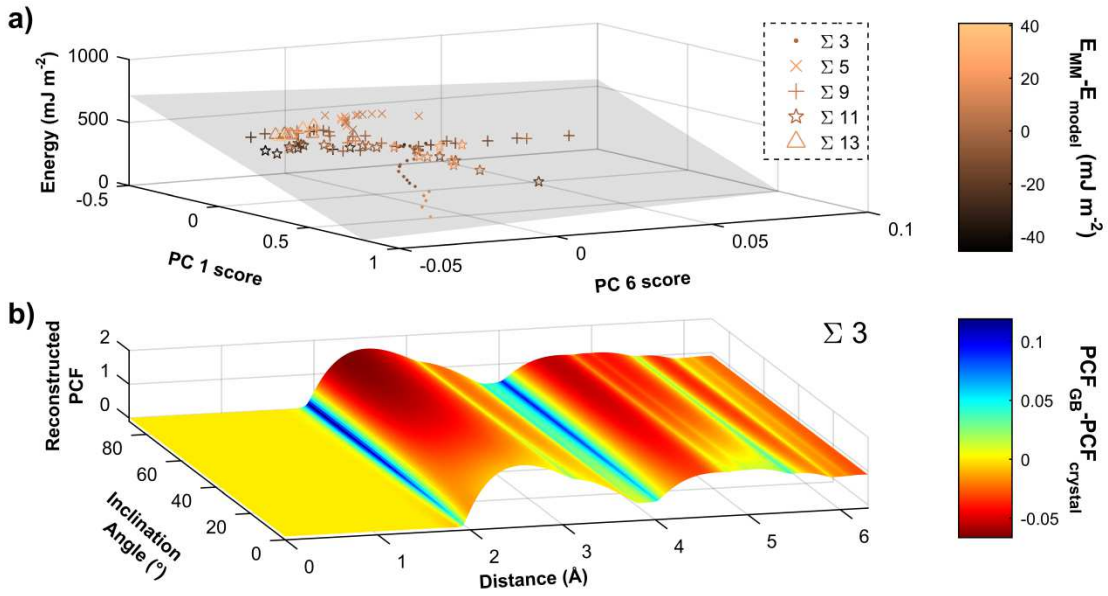

Figure 4.6.3 – Illustration of PSP linkages. (a) Structure-Property linkage: A plane representing the fitted regression model overlaid with the GB energy from simulation plotted against actual scores for PCs 1 and 6. The color scale represents the error of the regression model in mJ/m^2. (b) Process-Structure linkage: Continuous value of the predicted PCF as a function of inclination angle for a Σ3 ATGB . The color scale represents the deviation from the perfect crystal PCF.

The method outlined here provides a framework for efficiently extracting quantitative and transferable PSP linkages from molecular mechanics/dynamics simulations. Figure 4.6.3 illustrates the continuous nature of these linkages. The structure-property relationship illustrated in Figure 4.6.3a can predict the GB energy for

any Al ATGB with a reasonably similar structure to those in the model. The process-structure relationship can predict the structure itself as a function of θ and Σ (see Figure 4.6.3b). These linkages will aid in the coupling of complex GB boundary structures into multiscale models where hundreds or thousands of different GB structures may arise.

CHAPTER 5. PSP LINKAGES IN SYMMETRIC TILT GRAIN BOUNDARIES USING ASCA

5.1 Overview

The methods for establishing PSP linkages in GB simulations defined previously are further refined though the incorporation of ASCA (ANOVA single-component analysis), an extension of PCA that incorporates the additive decomposition of ANOVA to separately analyze variance from different known sources. In this work, 150 symmetric tilt GBs (STBGs) were analyzed with varying misorientation angles, axes, and micro-degrees of freedom in a manner consistent with two-way ANOVA. A 3-component structure-property regression model for predicting GB energy constructed using the first two sets of angle-PC scores and the first set of axis-PC scores had an average prediction error of 15.4 mJ/m^2. Regression models using a third order polynomial of the misorientation angle served as the structure-property models for predicting the first two angle-PC scores. These models had R^2 values of 0.95 and 0.93 and were robust against leave-one-out cross-validation (LOOCV).

5.2 Description of Data

A dataset of symmetric tilt GBs (STGBs) was generated by Srikanth Patala and Arash Banadaki. In this dataset, the misorientation axis of each GB was either [001], [011]. or [111]. The misorientation angles selected are enumerated in Table 5.2.1.

[001]	[011]	[111]
2.2466°	2.2505°	2.2551°
5.2649°	5.2807°	5.289°
8.2552°	8.2539°	8.2556°
11.235°	11.218°	11.241°
14.25°	14.226°	14.249°
17.231°	17.232°	17.236°
20.249°	20.257°	20.248°
23.223°	23.202°	23.225°
26.268°	26.261°	26.249°
29.242°	29.265°	29.255°

Table 5.2.1 – Misorientation angles simulated for each axis

Here, all angles are spaced 3±0.05° apart, and the range of corresponding angles across the different misorientation axes is less than 0.025°. They can therefore be treated as the same equally spaced angles to within a rounding error. For each combination of axis and angle, GB structures were generated corresponding to the different possible configurations of the micro-degrees of freedom, and 5 of these GBs were selected at random, with probabilities estimated by inputting the total energy of each GB (normalized GB Energy × cross-sectional area) into a Boltzmann distribution at 300 K. The GB energies of the replicates selected are depicted in Figure 5.2.1.

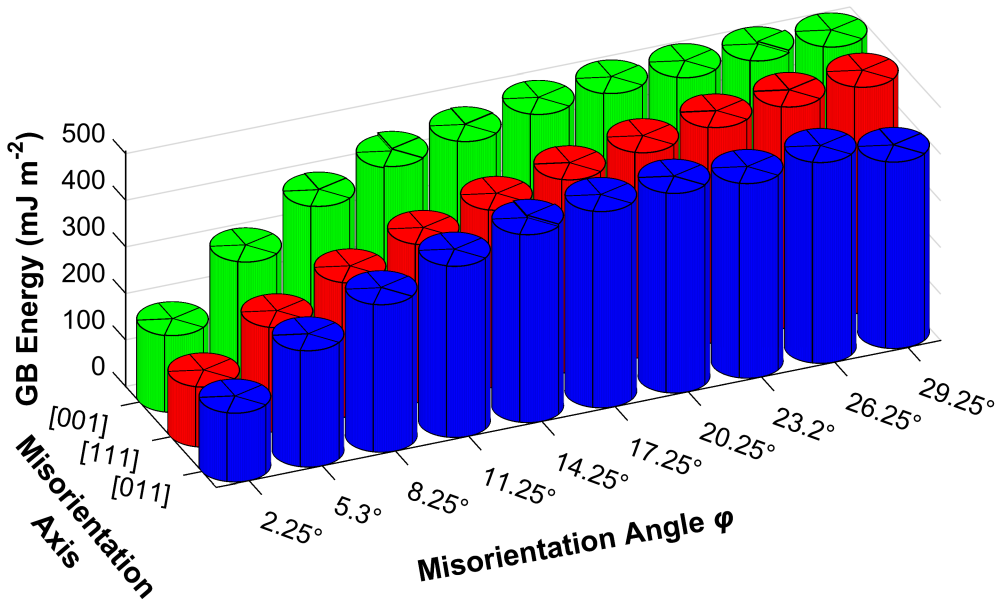

Figure 5.2.1 – GB Energies for each simulation included in the analysis.

5.3 Quantification of grain boundary structure

Grain boundary atoms were identified using the procedure outline in Section 2.2.2. Due to the highly organized nature of the variance structure of this dataset, no smoothing was required for the structure quantification, which was characterized with a traditional binned PCF using 95 bins from 0 to 6.3 Å. The PCF, thusly described, was calculated for each of the 150 GBs selected for analysis.

5.4 Low-Rank Model Construction

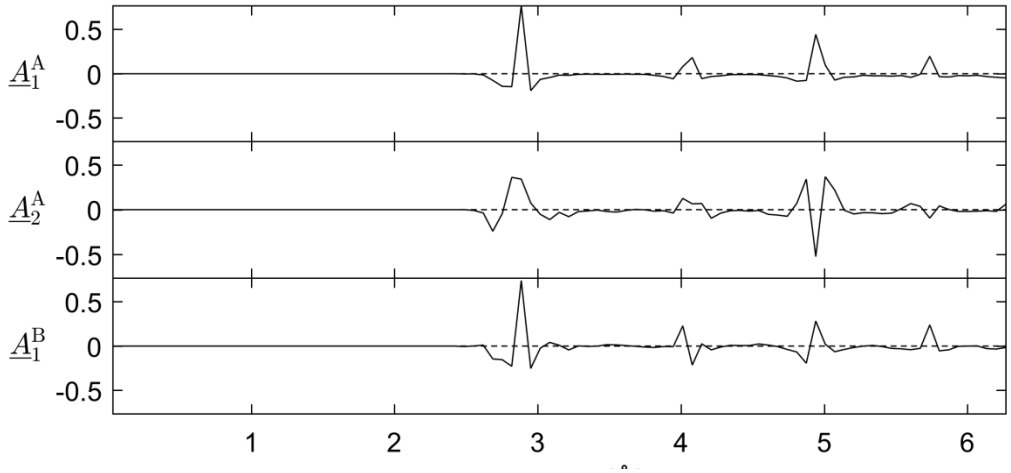

Figure 5.4.1 – Examples of ASCA eigenvectors.

The basis for the low-rank model established for this data set is ASCA (see Section 2.3.5). Here, misorientation angle served as factor A and misorientation axis served as factor B. A few of the corresponding ASCA eigenvectors are illustrated in Figure 5.4.1. The fraction of retained variance is illustrated in Figure 5.4.2 for a few sets of scores. From the ANOVA-style decomposition, it can be seen that structure differences corresponding to changes in misorientation angle account for 50.4% of the total variance. Changes in misorientation axis represent 33.1% of the variance with angle-axis interaction accounting for 15.3% of the variance and the remaining 1.2% corresponding to structure changes among the replicates within the same axis-angle grouping.

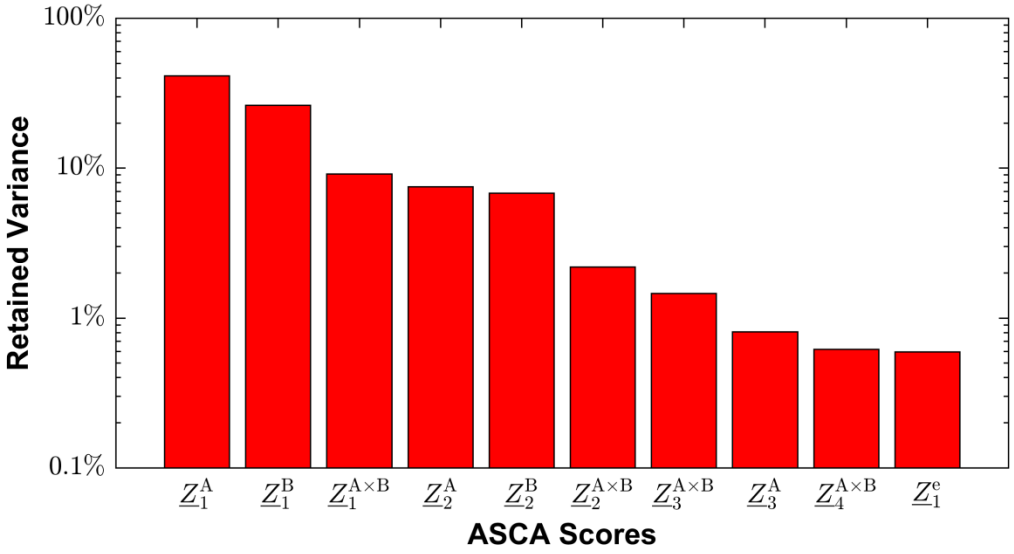

Figure 5.4.2 – Retained variance corresponding to the largest ASCA-PCs. Misorientation angle corresponds to factor A; misorientation axis serves as factor B.

Since the ASCA scores are orthogonal, the best *k*-component regression model consists of the sum of the *k* best 1-component models for the purposes of establishing a structure-property linkage. As such, separate regression models for predicting GB energy were constructed for each set of angle, axis, interaction, and replicate scores.

5.5 Results and Discussion

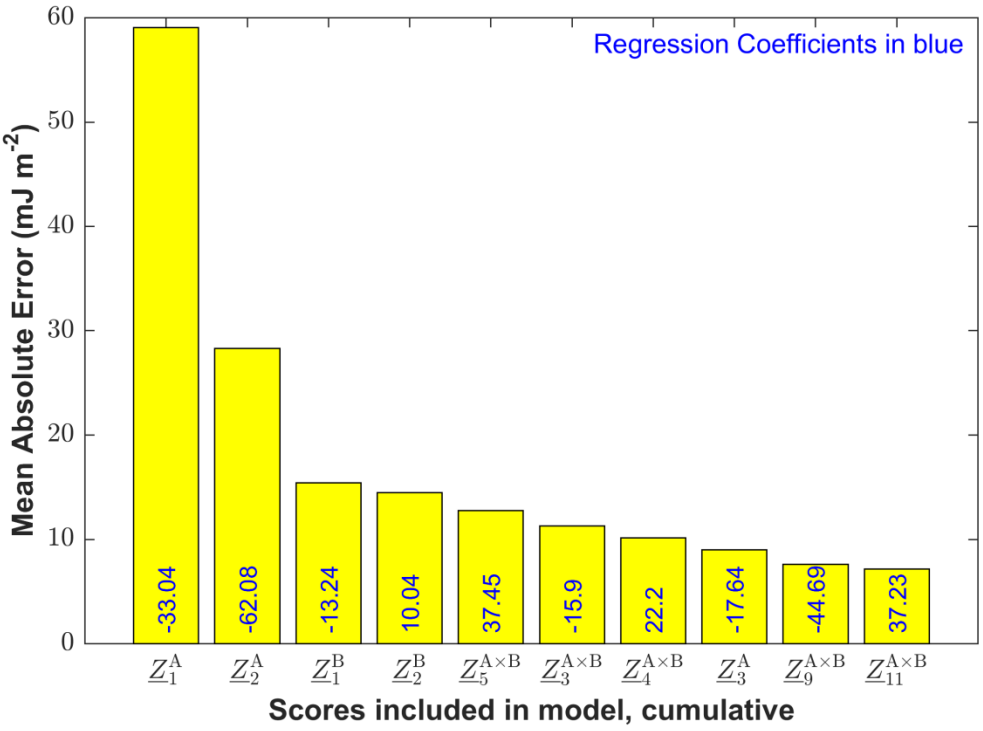

Figure 5.5.1 – Mean absolute errors resulting from the inclusion of each next-best set of scores in a linear regression model, with corresponding regression coefficients.

Figure 5.5.1 lists the best ASCA scores for predicting energy in descending order, along with the corresponding mean absolute error and corresponding regression coefficient. From here it can be seen that a 3-component model is optimal, consisting of angle-PC 1 and 2 scores, along with axis-PC 1 scores. Inclusion of additional PC scores yields a negligible marginal improvement in the model error. From the ordering of the scores in Figure 5.5.1, it can be seen that changes in misorientation angle have a greater influence on GB energy than changes in misorientation axis (consistent with the Read-Shockley model of low-angle GBs[85, 86]), and that both of these types of changes have a larger impact on GB energy than axis-angle interaction effects. Furthermore, structure changes among the replicates corresponding to each axis-angle pair have a negligible

effect on predicted GB energy. The parity plot of the 3-component model is depicted in Figure 5.5.2. From here, it can be seen that the model is more accurate for lower angles than for higher angles.

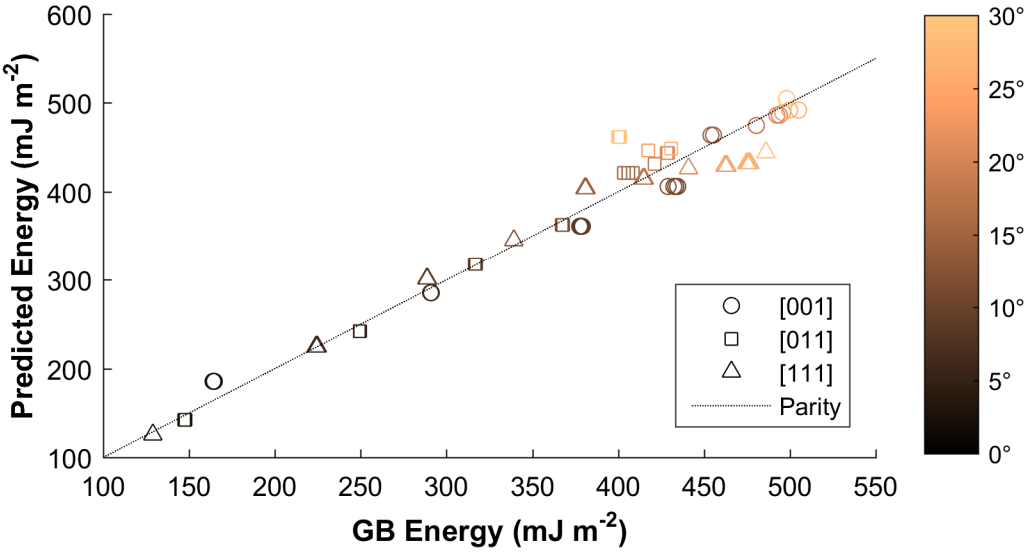

Figure 5.5.2 – Parity plot of 3-component linear regression model constructed from $\underline{Z}_1^{\mathbf{A}}$, $\underline{Z}_2^{\mathbf{A}}$, and $\underline{Z}_1^{\mathbf{B}}$.

The 3-component model is constructed from the first 2 sets of angle-PC scores and the first set of axis-PC score. The values of these scores are shown in Figure 5.5.3. From here, process-structure models regression models based on a third-order polynomial of the misorientation angle φ can be constructed for angle-PC scores included in the model.

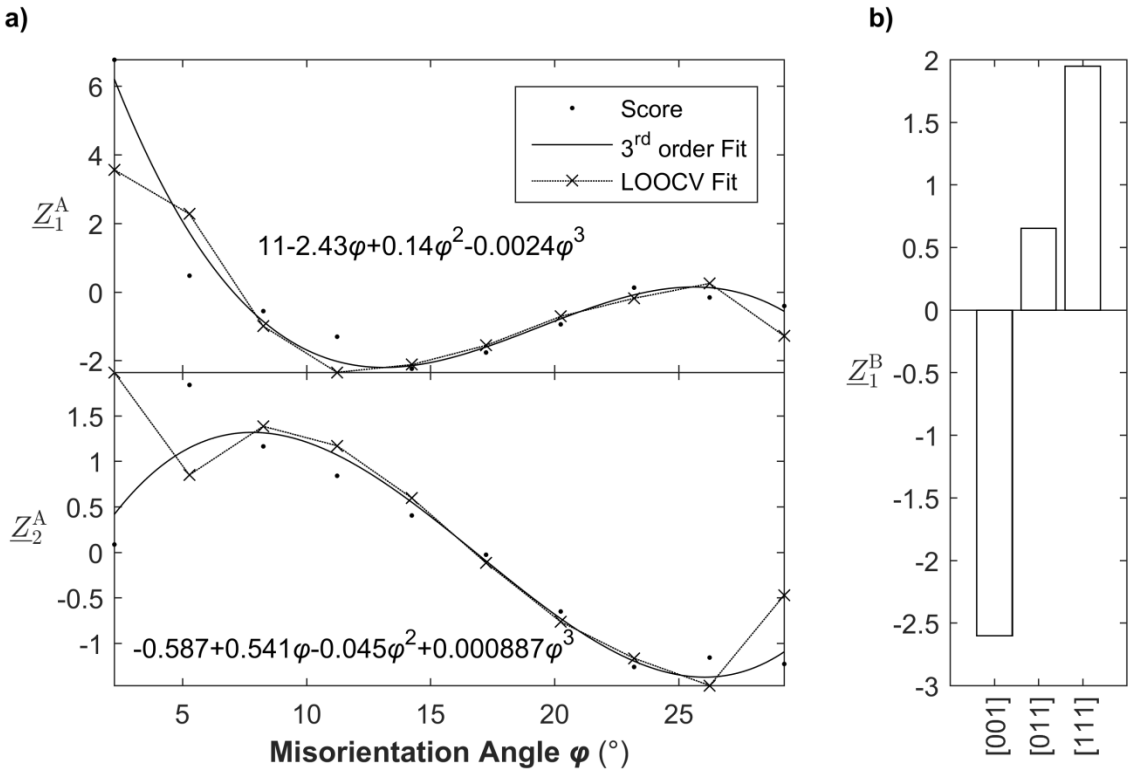

Figure 5.5.3 – Values of the ASCA scores associated with (a) misorientation angle, with corresponding structure-property model and (b) misorientation axis

The LOOCV error is small for most angles included in this study, with the exception of the highest and two lowest angles included in the study. This suggests that these structure-property models can be used to interpolate the angle-PC scores for misorientation angles toward the middle of the range of angles used here.

This study illustrates the benefits of ASCA over PCA for cases where there exists some prior knowledge of the sources of variance in a given data set. To obtain a structure-property model at least as good in terms of error as the 3-component regression model found here using traditional PCA regression, 7 PCs would need to be included in the model. Furthermore, process-structure models can be established that are valid for multiple misorientation axes. The R^2 value for the structure-property models for $\underline{Z}_1^{\mathrm{A}}$

and $\underline{Z}_2^{\mathrm{A}}$ are 0.95 and 0.93, respectively. In contrast, a 3$^{\mathrm{rd}}$-order polynomial of φ for predicting the PC 1 scores from traditional PCA has an R^2 of 0.53. The methods outlined here, coupled with those from CHAPTER 3, can further aid in the incorporation of GB structures into multiscale models.

CHAPTER 6. CONCLUSIONS

6.1 Relative Importance of Current Work

Perhaps the most important contribution described within this document is the method for calculating 2-point statistics for atomistic data outline in Section 2.1. 2-point statistics represent the primary structure characterization metric used within the multiscale modeling framework pioneered by the MINED Materials group; a computationally efficient method for calculating the 2-point statistics for atomistic data allows for the facile integration of atomic simulations into this framework. While defined here for the hard-sphere model, this method can be readily adapted for any meaningful spherically-symmetric descriptor of an image of an atom, such as a multivariate Gaussian distribution.

Furthermore, it was shown in CHAPTER 3 that atomistic 2-point statistics data can be used to categorize interatomic potentials based upon the results of simulations implemented with these potentials. Intelligent force field selection is a major barrier to more widespread usage of atomistic simulations in materials research, and this classification scheme can simplify the task of comparing the relative merits of different force fields.

The incorporation of ASCA, currently a little-used technique, into the data-driven materials modeling framework also holds great promise. In traditional PCA, associating the principal component eigenvectors with a particular source or sources of variance is a non-trivial task. In contrast, this is a quite simple matter in ASCA. As such, stronger

process-structure models can be constructed that link ASCA scores to the values of these known variance sources. Additionally, ASCA allows for more qualitative knowledge to be gleaned from structure-property models; a claim that a particular source of variance is more influential in determining a particular material property than another can be made based upon the relative predictive power of regression models constructed using the appropriately sourced ASCA scores as predictors.

The strength of the models constructed for the simulated grain boundary case studies described in CHAPTER 4 and CHAPTER 5 reinforces the notion that data-driven models operating under the PSP paradigm are a viable way to characterize material behavior regardless of length scale. Models such as these have the potential to improve the accuracy and efficiency of future simulations. The ability to rapidly predict a GB structure from a given set of simulation conditions using a process-structure model has the potential to reduce the number of computations required for future energy minimization simulations. To find the global energy minimum corresponding of a grain boundary corresponding to a given description of the macroscopic degrees of freedom, numerous structures based upon variations of the micro degrees of freedom must be investigated; the number of such structures might be reduced by restricting the potential structures examined to those sufficiently close to the structure predicted by the learning model. Additionally, GB energies predicted structure-property models have the potential to enhance mesoscale simulations of material properties influenced by GB energy, such as recrystallization, plasticity, and failure[34].

6.2 Future Work

6.2.1 Methodology

For the truncated 2-point statistics algorithm described in Section 2.1, the methodology for calculating and binning the displacements between atom centers in continuous space (described in Equation 46) may be improved. The simplest (but most computationally expensive) method for calculating these displacements prior to binning would involve using the modulo operator (see Equations 34 and 38) and then discarding the results that lie outside the required cutoff region. Rather than discarding long vectors, a method for only calculating displacements within the cutoff regime may be implemented. Possible candidate algorithms include range trees[87], octrees[87, 88], and local sensitivity hashing[89].

The method for calculating orthonormal eigenfunctions from smoothed functional PCA and ASCA (see Sections 2.3.7-2.3.8) may prove quite useful to the establishment of process-structure and structure-property linkages due to the introduction of a smoothing hyperparameter into the calculation of eigenvectors. The problem as is defined here for the case of one-dimensional functions, making it ideal for structural representations such as the PCF. Further refinement of the problem to define smoothing in multiple dimensions will allow for this technique to be implemented for structures characterized by 2-point statistics or higher-order PCFs (see Section 2.2.3).

6.2.2 Case Studies

All studies described here have focused on simulations involving only a single type of atom. A natural extension of this work would be to examine simulations with multiple types of atoms, such as aluminum and copper. For this sort of study, the structure would need to be characterized with multiple sets of 2-point statistics or PCFs to sufficiently characterize Al-Al, Cu-Cu, and Al-Cu displacements.

The methods employed to establish PSP linkages in GBS outline in Section 2.2.2 exploited the fact that force fields like the one used in these studies based on the Embedded Atom Model, a popular class of force fields for hard materials based exclusively on interatomic distances, to justify the use of PCF as the chosen structure metric; since the force-fields employed in the studies do not depend on the relative orientation of the atomic bonding structure in terms of angular positions, the structure metric need not retain information pertaining to orientation. However, force fields for soft material simulations such as AMBER[90], CHARMM[91] and MM2[92] incorporate dihedral and torsional angles of atomic bonds into the energy calculation. As such, the PCF would likely prove to be an insufficient structure metric in these cases. For soft materials, the n-point statistics of the atomic structure would be a potential satisfactory structure metric. Additionally, a higher-order PCF (described in Section 2.2.3) may be useful as it implicitly contains orientation information despite only characterizing neighbor distances.

The case study incorporating ASCA into process-structure and structure-property models (see CHAPTER 5) used a dataset of simulated STGBs arranged in a manner similar to a two-way ANOVA problem. To do this, the two factors selected to be varied were the misorientation angle and the misorientation axis, the latter of which was treated

as a categorical variable. As such, a meaningful process-structure model for predicting the ASCA scores associated with changes in the misorientation axis could not be constructed. However, the misorientation axis need not be treated as a categorical variable; the normalized vector defining the axis can be described in terms of the azimuthal and elevation angles, or an equivalent 2-degree-of-freedom representation. As such, it is possible to follow up the work done in CHAPTER 5 with a study using a new STGB dataset arranged in the manner of a three-way ANOVA problem. Here, the three factors would be the misorientation angle, the misorientation axis azimuthal angle, and the misorientation axis elevation angle. For this problem, the factor levels should be as evenly spaced as possible, and the angles corresponding to the same factor level should be as close to equal as possible, given the constraints of rational numbers.

APPENDIX A. APPLICATION OF DATA SCIENCE TOOLS TO QUANTIFY AND DISTINGUISH BETWEEN STRUCTURES AND MODELS IN MOLECULAR DYNAMICS DATASETS

A.1 Abstract

Structure quantification is key to successful mining and extraction of core materials knowledge from both multiscale simulations as well as multiscale experiments. The main challenge stems from the need to transform the inherently high dimensional representations demanded by the rich hierarchical material structure into useful, high value, low dimensional representations. In this paper, we develop and demonstrate the merits of a data-driven approach for addressing this challenge at the atomic scale. The approach presented here is built on prior successes demonstrated for mesoscale representations of material internal structure, and involves three main steps: (i) digital representation of the material structure, (ii) extraction of a comprehensive set of structure measures using the framework of *n*-point spatial correlations, and (iii) identification of data-driven low dimensional measures using principal component analyses. These novel protocols, applied on an ensemble of structure datasets output from molecular dynamics (MD) simulations, have successfully classified the datasets based on several model input parameters such as the interatomic potential and the temperature used in the MD simulations.

Keywords: multiscale modeling, principal component analysis, and molecular dynamics

A.2 Introduction

Multiscale modeling [1-4] has been identified as the most promising avenue for accelerating the design, development, and deployment of new/improved materials in emerging technologies [93-97]. A number of recently announced national research strategic initiatives (e.g., [93, 96, 97]) are being built on the premise that an increased use of multiscale materials modeling can dramatically reduce the need for extensive (and often expensive) experimentation that dominates the current materials development efforts. However, the main factors impeding the highly desired increased utilization of multiscale modeling can be collected into three groups [6]: (i) Model Maturity (i.e., the accuracy and reliability of available models), (ii) Model Interoperability (i.e., ability of the models covering multiple scales and physics to be strung together to work seamlessly), and (iii) Model Inversion (i.e., ability to address high value problems of interest in materials and process design that target improvements in specific performance needs). It should be noted that tremendous progress has indeed been made in being able to numerically simulate a broad range of materials phenomena using sophisticated physics-based modeling approaches [1-4, 7-17]. However, it is essential to address the main impediments described above, if we are to realize the full benefits from these modeling approaches in advanced materials development efforts.

Modern data science tools and concepts offer a promising new avenue for addressing most of the impediments described above. Data science [23-26] is mainly focused on extracting high value information (might be labeled as knowledge or wisdom) from all available data (generated by either experiments or computations). This emerging cross-disciplinary field is being built on the foundations of statistical sciences,

computational sciences, systems theory, and applied mathematics, and is envisioned to have a broad range of potential applications. Indeed, data science has already enjoyed many remarkable successes in disparate application domains, including recommendation systems (e.g., Amazon [98]), personal informatics (e.g., [99]), drug discovery (e.g., [100]), decision systems (e.g., [101]), and healthcare (e.g., [102]). At its core, data science is comprised of two primary components. The first component can be broadly identified as Data Management and includes robust and reliable storage, aggregation, archival, retrieval, and sharing protocols for all kinds of data (potentially generated in the broadest variety of formats possible). The second component (more pertinent to the present discussion) centers around Data Analytics, and is aimed at mining hidden (embedded) high value knowledge or understanding from large collections of data.

In the context of advanced materials development efforts, the central goal of Data Analytics is the extraction of robust and reliable process-structure-property (PSP) linkages that capture quantitatively the roles of different unit manufacturing (or processing) steps on the salient measures of the material hierarchical structure that in turn control the properties of interest (or performance characteristics desired in service). In this regard, it is extremely important to cast the desired PSP linkages in computationally efficient forms that allow direct integration into the tools typically employed by practitioners in the product design and manufacturing fields. In other words, the PSP linkages of interest are not likely to be employed in the forms developed in the advanced numerical tools [1-4] or the sophisticated homogenization theories [103-108], but more likely in the reduced-order forms (also called surrogate models or metamodels) that allow practical solutions to inverse problems of materials and process design. In recent years, a

data-centered framework has emerged for capturing highly accurate PSP linkages relevant to a broad range of materials phenomena [109-120]. Almost all of the applications demonstrated so far have focused on meso-length scales in the material internal structure. For example, the relationship of mesoscale porous structures on effective transport properties has been investigated [33, 109, 121-123]. In this paper, we extend this prior framework to atomic-scale molecular dynamics (MD) datasets and demonstrate its viability as a tool for improved hierarchical modeling and as a means to characterize and distinguish between datasets used in atomistic simulations. Indeed, our goal is to use the same structure quantification techniques at the atomic scale as those used previously at the mesoscale. Consequently, the approach presented here paves the way for the development of an universal approach for the rigorous quantification of the material structure at multiple hierarchical length/structure scales.

A distinctive feature of the materials data science approach presented here is its focus on a rigorous, statistical, quantification of the material structure and its usage in arriving at PSP linkages. The underlying hypothesis in such an approach is that only a sufficiently comprehensive description of the material structure can facilitate the capture of robust and reliable PSP linkages (e.g., [6, 110, 124-126]). The central challenge, therefore, lies in the quantification of the material internal structure. A complete and rigorous description of the material internal (hierarchical) structure can be very complex, demanding very high dimensional representations. This challenge is readily appreciated when one recognizes the need to include not only the details of an idealized structure in the materials of interest, but also the inherent defects (including disorder) and their spatial distribution in the structure. For example, most materials being explored for structural

applications exhibit multiphase polycrystalline microstructures at the mesoscale [127-130]. A rigorous description of such material structures should include quantification of the spatial distributions of the chemical composition, thermodynamic phases, crystal lattice orientations and various hierarchical defect populations (e.g., point defects, dislocations, grain boundaries, phase boundaries, pores, microcracks). Fortunately, the field of materials science and engineering has already taught us that only certain salient features of the material internal structure dominate the macroscale performance characteristics of interest for any selected application. Therefore, the main challenge in the development of materials with improved/enhanced properties reduces to identifying and tracking only the salient microstructure features that are important to a specific engineering or technology application. In general, these salient features of the material structure are not known a priori, and need to be identified from an extremely large list of potential measures. This is precisely where a data-driven approach offers many advantages. In a data-driven approach, the decision on exactly what constitutes the set of important salient features is not taken in a static manner – instead it is taken objectively based on the actual available data. It is continuously refined as more data becomes available.

A major goal of this work was to test whether the methods previously developed for mesoscale structure quantification could be applied to atomistic “samples” produced by MD simulations. In particular, our goal was to explore if these methods can objectively distinguish between atomic configurations in a way that would support multiscale modeling. In this work, the results using different interatomic potentials (models of energies and forces between atoms) were considered a surrogate for different processing

methods. It is important to distinguish objectively between results generated by different models and/or under different simulation conditions. Another important factor is that, by making use of robust global characterization methods, it is possible to establish greater confidence in the multiscale use of the results from classical MD simulations.

The structure quantification approach presented in this paper, and applied rigorously to MD datasets for the first time, comprises three essential steps. In the first step, the output from the MD simulations presented as expected positions of the atom centers, is transformed into a digital (uniformly tessellated) structure. In the second step, the digital representation of the material structure is quantified using the framework of *n*-point spatial correlations (or *n*-point statistics) [110, 127, 131-134]. Although a number of other ad-hoc measures of material structure are possible, only the *n*-point spatial correlations provide the most complete set of measures that are naturally organized by increasing amounts of structure information. For example, the most basic of the *n*-point statistics are the 1-point statistics, and they reflect the probability of finding a specific discrete local state of interest at any randomly selected single point (or voxel) in the material structure. In other words, they essentially capture the information on volume fractions of the various distinct local states present in the material system. The next higher level of structure information is contained in the 2-point statistics, denoted $f_r^{hh'}$, which capture the probability of finding discrete local states h and h' at the tail and head, respectively, of a prescribed vector r randomly placed into the microstructure. This idea is closely related to the commonly used concept of pair correlation functions [73] that reflect, for a selected or representative atom, the probability of finding atoms (generically or of a given type) as a function of radial distance. The main difference between the pair

correlation functions and the 2-point correlation functions is that the latter capture the directional dependence, i.e., the difference between the points examined is expressed as a vector and not just a simple scalar distance.

The third and final step of structure quantification involves the objective identification of reduced-order representations of the structure using techniques such as the principal component analysis (PCA) [110, 135]. PCA provides a linear transformation of high dimensional data in a new orthogonal frame where the axes are ordered according to the observed variance among the elements of the dataset. Consequently, a truncated PCA representation provides an objective (data-driven) reduced-order representation of the original data. It is emphasized here that although PCA dimensionality reduction techniques have been explored in materials problems in prior literature [12, 136], they have only recently been employed on 2-point spatial correlations of microstructure in attempts to successfully extract high fidelity PSP linkages [109, 110, 135, 137]. The main contribution of this paper is a demonstration of the application of these computational toolsets on MD datasets, and to compare and contrast the results with those obtained using the simpler structure measures used currently. Although further development of the ideas presented here is needed before they can be broadly adopted, this work demonstrates the viability and advantages of employing spatial statistics and PCA protocols on the MD datasets.

A.3 Background: Spatial Correlations

As noted earlier, structure quantification is central to the extraction of transferrable materials knowledge needed in multiscale materials modeling efforts. A digital signal

representation of the material structure serves as a natural starting point for the ensuing discussion. In particular, it has been proposed to represent the discretized material internal structure as m_s^h [138], which denotes the probability that a specified spatial bin (or voxel) indexed by s is physically occupied by a potential local state indexed by h. Since the values of m are bounded between zero and one (in many cases it can be just binary [138]), it produces a generalized representation for a broad range of materials systems at different length/structure scales. The information on the different length scales is encoded into the properties associated with the spatial bins, while the information on the local state of the material (e.g., chemical composition, phase identifiers, order parameters, tensorial representations of different defect configurations of interest) is encoded into the properties associated with the bins in the local state space. The digital signal representation of structure offers many computational advantages in a broad range of materials data transformations and knowledge extractions [110, 113-116, 123, 127, 129, 133, 134, 137, 139-143].

The material structure representation described above is particularly well suited for the computations of spatial correlations (i.e., information on the relative placement of local states in the material structure) [110, 127, 131-134]. Based on the earlier definitions, the 2-point spatial correlations (or 2-point statistics) can be mathematically expressed as [133, 134]

$$f_r^{hh'} = \frac{1}{S_r}\sum_{s=1}^{S_r} m_s^h m_{s+r}^{h'} \tag{150}$$

where r indexes the bins in the space of vectors (generally the same binning scheme as that was used for the spatial domain). In Eq. 150, S_r denotes the number of spatial bins for which the bins indexed s and $s + r$ both lie within the spatial domain of the material structure instantiation being studied. If assumptions of periodicity of the material structure are invoked (as routinely done in MD simulations), then $S_r = S$, where S is the total number of spatial bins in the microstructure instantiation. It is also pointed out that computationally efficient schemes for computing the spatial correlations using Discrete Fourier Transforms (DFTs) have been developed and utilized successfully [133, 134].

For most structural material systems of interest in advanced technologies, the set of *n*-point statistics is an extremely large unwieldy set even for *n* = 2. Rigorous analysis of these datasets is only possible with the application of data science tools. For example, it was recently demonstrated that techniques such as principal component analysis (PCA) [144-146], can be used to obtain objective low dimensional representations of the 2-point statistics [110, 135]. PCA provides a linear transformation of high dimensional data in a new orthogonal frame where the axes are ordered according to the observed variance among the elements of the dataset. Consequently, a truncated PCA representation provides an objective (data-driven) reduced-order representation of the original data. It is emphasized here that although PCA dimensionality reduction techniques have been explored in materials [12, 136] and biology [42, 147-150] problems in prior literature, they have only recently been employed on 2-point spatial correlations of microstructure in attempts to successfully extract high fidelity structure-property linkages [109, 110, 135, 137].

As an example, let $\{f_r | r = 1,2, \dots, R\}$ denote the truncated set of independent 2-point statistics [133] of interest in a specific application. Let $i = 1,2, \dots, I$ enumerate the elements of an ensemble of material structures being studied. It is generally expected that $I \leq R$. In such situations, PCA identifies a maximum of $(I - 1)$ orthogonal directions in the R-dimensional space that are arranged by decreasing levels of variance in the given ensemble of structures. Mathematically, the PCA representation of any member of the selected ensemble (of structures), labeled by superscript (k), can be expressed as

$$f_r^{(k)} = \sum_{i=1}^{\min((I-1),R)} \alpha_i^{(k)} \varphi_{ir} + \bar{f}_r \tag{151}$$

where $\bar{f}_r$ is simply the averaged 2-point statistics for the entire ensemble, and $\alpha_i^{(k)}$ (referred as PC weights) provide an objective representation of the $(k)^{th}$ structure in the new orthogonal reference frame identified by φ_{ir} (from PCA). Another important output from the PCA is the significance of each principal component, b_i, obtained in the eigenvalue decomposition performed as a part of the PCA [144-146]. The values of b_i provide important measures of the inherent variance among the members of the ensemble of structures [135]. More importantly, by retaining only the components associated with the most significant eigenvalues, it is often possible to obtain an objective reduced-order representation of the structure with only a handful of parameters. Mathematically, this reduced-order representation can be expressed as

$$f_r^{(k)} \approx \sum_{i=1}^{R^*} \alpha_i^{(k)} \varphi_{ir} + \bar{f}_r \tag{152}$$

where $R^* \ll \min\big((I-1), R\big)$. Selection of R^* will depend on the specific properties that need to be correlated to the structure metrics. Note also that the concepts described above can be easily extended to include higher-order statistics of the structure (e.g., 3-point spatial correlations). The PCA representations of the *n*-point statistics have been successfully used in automated and efficient classification of various ensembles of structures [110, 137].

In most prior examples presented to date in literature, the local state was defined at the continuum scale and identified as a specific thermodynamic phase found in the micrograph. However, the same methodology can be applied to material structures at other length scales. In a recent paper, this approach was successfully applied to quantify the semi-crystalline polymer structure datasets produced by molecular dynamics (MD) simulations [151].

A.4 Extension of Spatial Correlations to MD Datasets

One challenge of applying 2-point statistics to atomistic configuration datasets is the subjective choice of how to transform the discrete set of atomic points into a regular three-dimensional (3-D) grid of voxels. This choice is likely to be driven by the nature of the application. For example, in simulations encompassing a relatively large number of atoms, it may be preferable for a single voxel to encompass multiple atoms and the local state in each voxel is defined by measures such as the density or the mean orientation of the enclosed atoms (e.g., [151]). Alternatively, it may be preferable to quantify structural variations at the atomic scale, in which case the voxel size should be selected to be smaller than the atomic radius; we will focus our discussion here to these cases.

As a proxy for more complex atomic structures, we here consider MD simulations of atomic volumes with a single chemical species as a function of temperature. These simulations represent relatively simple MD calculations that are being used as part of the NIST Interatomic Potentials Repository project to help establish a set of reference calculations to help researchers select interatomic potentials (models of how the atoms interact, also called force fields) that are most appropriate for a given application [152]. Except for choice of interatomic potential, the methodology is kept fixed for every simulation, which is: (i) determine the 0 K equilibrium FCC lattice constant via a molecular statics simulation, (ii) create a 10 x 10 x 10 face centered cubic (FCC) unit cell (4000 atoms) using the equilibrium lattice constant, (iii) create a uniform distribution of atomic velocities at the desired simulation temperature, and (iv) perform an isothermal-isobaric (NPT) simulation at the desired temperature for 2,000,000 time steps using a 1 fs time step. Data analysis described here takes place within the final 1,000,000 time steps. Instantaneous coordinates were recorded every 50,000 fs, and these were used in the analysis presented here. Average reported pressures, volumes, temperatures, energies, etc., reach steady state well within that equilibration time for all simulations. The long simulations were done (instead of shorter ones that may have been adequate), primarily for two reasons. The first was to minimize the chance of a particular trajectory not being in equilibrium while running the same duration for all simulations (to make comparisons more robust). The second was to allow more time for first-order phase transitions to occur to thermodynamically favorable states. While this is not an issue for low homologous temperatures (T/T_M), it is more significant near the melting temperature of the interatomic potential where phase transitions (melting) were observed for several of

the interatomic potentials. Melting is identified by local structural disorder and a significant increase in atomic volume. The python scripts used to generate the simulations and the data itself are available on the NIST Interatomic Potentials Repository site (http://www.ctcms.nist.gov/potentials). While calculations have been performed for a number of different interatomic potentials defining elemental interactions for Al, Ni, Cu, Ag, and Au, here we are focusing on just the Al results. The interatomic potentials included in this study are summarized in Table 1, along with the appropriate references [60-79].

It is important to note that these calculations include some simulations well outside the intended usage of the interatomic potentials (e.g., using the pure elements of a potential only fit for use with compounds and thus they may not give the most accurate values for single-element atomic volumes). However, users often use interatomic potentials well outside the range of where they were fit, and it is important to understand how that choice affects the answers obtained. This is discussed in much more detail in Refs. [152-154]. In this work, several interatomic potentials have melting temperatures for pure aluminum that are significantly lower than the experimental value of 933 K, which will be discussed in more detail later.

Figure 3.2.1 shows a MD simulation dataset typical of those included in this study. In this dataset, the center positions of the atoms were taken directly from the results of the MD simulations (as instantaneous coordinates) and a sphere of radius $a = 1.18$ Å was constructed around the center to denote the atom. The entire volumetric domain used in the simulation was then discretized into a uniform grid and the material structure was converted to a simple digital signal, denoted as m_s^h (as introduced earlier). In this

notation, the local state descriptor, h, was allowed only two values: $h = 1$ was used to refer to the atomic species and $h = 0$ was used to refer to the empty space between the atomic species. As mentioned earlier, s serves as an index for the spatial bin. For 3-D space, it is convenient to think of s as an integer array, i.e. $s = \{s_1, s_2, s_3\}$, with each s_i taking only integer values. The level of discretization employed is typically a variable parameter in the data-driven explorations. In the present study, based on a few trials we established a spatial bin size of approximately 0.252 Å $= 0.214a$ since further refinement did not influence the computed spatial correlations in any significant manner. The value assigned to m_s^h denotes the volume fraction of local state h found in the spatial bin s. Although, in principle, the value of m_s^h can range between zero and one, we have only allowed this variable to take either the value zero or one in this study; such structures have been referred as eigen structures in prior literature [138]. More specifically, if the distance between the center of a given voxel and the center of the voxel containing the coordinates of the atom center is less than or equal to the radius, that voxel is assigned a value of one (i.e., the voxel is included in the atom). For eigen microstructures, f_0^{11} would actually be the volume fraction occupied by the atomic species in the total volume being studied. Furthermore, since there are only two local states in the datasets considered here, only one autocorrelation is enough to capture all of the non-redundant 2-point spatial correlations [124, 131, 133, 141, 155]. In this paper, we will therefore only focus on f_r^{11}, and simply refer to these as f_r.

Next, we discuss the computation of f_r from m_s^1. A specific challenge encountered arises from the fact that the overall simulation volume in the MD results is not kept constant. In other words, results from different potentials or even different

snapshots from a single potential are expected to result in different simulation volumes. Since we have fixed the spatial bin size (described above), this would lead to fractional voxels at the edges of the volume. Furthermore, since the MD simulations conducted for this study have employed periodic boundary conditions, we wish to rigorously account for these boundary conditions in computing the spatial correlations. The strategy devised and employed in this study, to address the considerations described above, consisted of the following steps: (i) The microstructure signal, m_s^1, is expanded by employing the same periodic boundary assumptions that were utilized in the MD simulations. As an example, this expansion is shown in Figure 3.3.1 for a representative 2-D section through the simulation volume in Figure 3.2.1. For this example, the domain volume size is increased from L_o=40.5 Å to L_e=73.08 Å (in each of the three dimensions). Note that this expansion serves two purposes: (a) While the initial volume size (output from the MD simulation) is unlikely to be an exact integer multiplier of the selected spatial bin size, the size of expanded region is selected to ensure that it is indeed an exact multiplier of the spatial bin size (this feature is essential to allow the use of DFT algorithms). (b) The increase in size is needed to allow the placement of all vectors of interest in computing the spatial correlations (the tails of the vectors of interest will lie within the original volume, but the heads of these vectors may lie in the expanded volume). For all the MD volumes included in the study, the expansion size was selected to include all vectors up to a size of 59 spatial bins (this number was selected after a few trials and noting that vectors larger than this do not carry any additional salient information in the computed 2-point statistics for the volumes studies here); the corresponding number of statistics will be 119^3 (59 positive, 59 negative, and the zero vector components in each of the three

dimensions). Discretization using finer grids was seen to have a negligible effect on the clustering (i.e., classification) of interest for the present study (visualized later as dendrograms; cf. Figure 3.5.2). It is important to note that the discretization level is an important parameter of the protocols described here, and has to be adjusted suitably for different studies. (ii) A second microstructure signal $\widetilde{m}_s^1$ of the same extent as m_s^1 is created by copying the values of m_s^1 for all of the spatial bins corresponding to atoms whose centers fit inside the original volume (of size L_o) and assigning zeros for the rest of the spatial bins (also shown in Figure 3.3.1). The number of spatial bins copied from the original volume is denoted as S_r. (iii) The 2-point spatial correlations of interest are computed as the convolution of the two microstructure signals, m_s^1 and $\widetilde{m}_s^1$ (i.e., using these instead of m_s^h and $m_s^{h'}$ in Eq. 150), truncated to include only vectors whose 3-D components are smaller than R.

Figure 3.3.2(a)-(c) presents selected 2-D sections of the 3-D contour plots of 2-point spatial correlations (these are visualized as the contours of the values of f_r in the 3-D vector space of r, with $r = (0,0,0)$ at the center of the plot). The sections shown in this figure depict, as expected, a roughly periodic pattern consistent with the crystalline structure reflected in the spatial positioning of the atoms in the actual volumetric domain analyzed by the MD simulations (shown in Figure 3.2.1). It is important to recognize that the f_r values plotted in Figure 3.3.2(a)-(c) are actually statistics denoting the probability of finding two voxels separated by the vector r and occupied by the atomic species. As a reference, the reader might take note that in a perfectly disordered (i.e., random) spatial distribution of local states (not shown), the 2-point spatial correlations show a single spike at the center (for $r = (0,0,0)$) and then immediately asymptote to an uniform value

as one moves away from the center. The reader should also note that the value of $f_{(0,0,0)}$ at the center of these plots corresponds to the atomic volume fraction.

Figure 3.3.2(d) presents the more familiar pair correlation function (PCF) used extensively in literature for quantifying the material structure in the MD simulation results. As one might infer, the peaks in the PCF plot correspond to suitably integrated (and normalized) values of the 3-D 2-point spatial correlations over the orientation variables defining the vector r. In other words, PCF is expressed only as a function of the magnitude of r, while the 2-point spatial correlations retain explicitly the dependence on both magnitude and direction of r.

A.5 Application of Spatial Correlations to MD Datasets

Figure 3.5.1 presents a classification of the MD simulation datasets in the PCA space (following the protocols described earlier) for the MD simulated atomic structures at 300K and 900K, respectively, using the 19 potentials selected for this study. For each potential, the study included twenty atomic structures (taken at different times in the simulation after reaching an equilibrium state). Therefore, a total of 380 atomic structures were included in this analysis at each simulation temperature. Each data point in Figure 3.5.1(a) and (b) represents the first three PC scores (or weights) for each MD simulated atomic structure included in the analyses. The computation of course provides many more dimensions (or PC scores), but it also indicated that these three PC scores account for 99.8 % of the important differences in the entire ensemble of atomic structures included in the study. This massive degree of dimensionality reduction is fully consistent with the prior experience involving mesoscale systems. In this regard, it is also satisfying

to note that the range of the PC scores is systematically decreasing for the higher-ranked PC scores (for example, the range for PC1 was about -90 to about 20, whereas the range for PC2 was about -15 to about 15), further confirming that the higher-ranked PC scores are indeed less important in capturing the salient features of the structures included in the ensemble.

Keeping in mind that the PCA representation in Figure 3 denotes a dimensionality reduction from 119^3=1685159 to just three, it is indeed remarkable that this representation effectively captures both the intra-class and the inter-class variations within the entire ensemble. This result is even more remarkable when one notes that this classification was performed in a completely unsupervised manner. In other words, the PCA computation was not informed in any way about the different potentials used in the MD simulations in producing the atomic structures included in the study. This is a clear testament to the power of the 2-point spatial correlations and principal component analyses in capturing the salient features of the material structure in a rigorous stochastic framework. It is also very satisfying to note that the intra-class variance (reflected in the size of the cluster associated with each potential) in the simulated structures is significantly smaller than the inter-class variance. Moreover, the intra-class variance seems to be of roughly the same order of magnitude for all the different potentials included in this study, and is slightly higher for the datasets produced at the higher simulation temperature. All of these observations are consistent with expectations, and provide strong support to our claim that the protocols used in this study produce high value, low dimensional, measures of the material structure.

An effective tool for visualizing distances in high-dimensional spaces is a dendrogram, which depicts the hierarchy of the distances between the data points. Figure 3.5.2(a) and (b) depict the inter-class distances (between the cluster-means) as dendrograms for the same dataset that was depicted in Figure 3.5.1. Broadly, the PCA has identified the following clustering of potentials based on the differences in the structures produced by the MD simulations: the first group corresponds to the force fields referenced in [61, 64, 65, 69, 72], the second group corresponds to the force fields referenced in [60, 62, 66-68, 71, 73, 77, 78], including both force fields referenced in [71]. The four force fields referenced in [63, 70, 76] and [74, 75] are distinctly far away from the two groups identified above. The groupings of these results will be discussed in more detail in a later section. Here we reiterate that interatomic potentials are fit with different types of reference data and optimized for particular applications. Potentials fit for particular compounds, e.g., the B2 phase in Ni-Al, may not be the best option for treating the full Ni-Al phase diagram, though they may be the best available for the intended application.

Additional insights from the analysis presented here can be obtained from the plots of the PCs obtained in the analysis described above. Plots of $\bar{f}_r$ and φ_{ir} (for different values of i; see Eq. 152 are presented in Figure 3.5.3(a)-(d). As with the plots shown in Figure 3.3.2(a)-(c), r indexes the discretization of the vector space used in defining the 2-point spatial correlations. The plots of $\bar{f}_r$ (Figure 3.5.3(a)) simply reflect the averaged auto-correlations for the entire ensemble of atomic structures included in the study. As expected, the averaged auto-correlation reflects an arrangement of the atoms on a highly periodic lattice. One can judge the degree of periodicity by comparing intensities

of the different peaks in these plots with the intensity of the center peak. For a perfectly periodic arrangement, the peak intensity will be the same for all peaks in the entire plot. As the arrangement becomes less periodic, the peak intensities drop systematically as one moves away from the center peak. As mentioned earlier, for a random arrangement, this drop in the peak intensity will be rather abrupt. In the present study, we will see a more significant drop in the peak intensities for the atomic structures simulated at higher temperatures (described later) compared to the ones simulated at lower temperature.

The plots of φ_{ir} in Figure 3.5.3(b)-(d) reflect a prioritized set of orthogonal deviations from the averaged autocorrelation. In other words, φ_{1r} reflects the most dominant deviation, φ_{2r} is the next most dominant deviation, and so on. Note the difference in signs between the red and black peaks in these plots. Consequently, a combination of closely placed pair of red and black spots on these plots reflects shift of the peak from its position in the ensemble average. The overall plot of φ_{1r} therefore captures systematic shifts in the interatomic distances between any selected atom and its neighbors, with the shifts being higher for far away neighbors compared to those that are nearby. Therefore, φ_{1r} appears to capture well the overall volume differences among the snapshots of the atomic structure. In the most general case, each of the φ_{ir} captures a certain scaled deviation in the intensities of all of the statistics included in the PCA analyses. Because of the large number of the statistics included in the PCA (each structure is characterized by 1,685,159 2-point statistics), it is often very difficult to assign a simple interpretation for what detail of the structure is captured by each individual φ_{ir}. It should also be noted from Figure 3.5.3 that the structure detail captured by the different φ_{ir} exhibit different levels and types of directional dependence.

As implied in Eq. 152, one can construct the autocorrelation of any specific atomic structure included in the study by starting with the averaged autocorrelation and adding weighted contributions from each of the principal components. These weights are precisely the weights depicted in the low dimensional PCA representations of Figure 3.5.1(a) and (b). It should be noted that such a reconstruction typically involves a truncation error when the higher-order principal components are ignored. However, since PCA provides a prioritized list of principal components, one can make the decision on an appropriate truncation level for a specific study in a very objective manner.

Figure 3.3.3(a) and (b) compare the 2-point statistics for the atomic structures predicted by one force field at two temperatures, respectively: 300 K and 900 K. As mentioned earlier, one of the salient differences in these plots is in the rate of decay of the peak intensities as one moves from the center peak, indicating the existence of a higher level of disorder (thermal noise) in the atomic structure at the higher temperature. It should be noted that this is a statistically rigorous evaluation of the difference in the atomic structures at the two temperatures. There is also a difference in the lattice parameter at the two temperatures, which can be easily inferred by looking closely at the positions of the peaks (with respect to the center) in the plots presented in Figure 3.3.3.

It is also instructive to examine the variation of the PC scores as a function of temperature for the different force fields. This is shown in Figure 3.5.4(a) and (b) after performing a PCA on all of the averaged 2-point statistics for each force field at each simulation temperature. Of particular interest are the four force fields corresponding to References [63, 70, 76] and [74, 75] , which show significantly different behavior compared to the rest of the data sets. Indeed, as shown in Figure 3.5.5, this difference in

the predicted results from these four force fields is also evident in the plots of the averaged atomic volume. The force field used in Ref. [76] was strongly weighted to reproduce the properties of B2-NiAl, which may explain its poor behavior for pure aluminum. The other three interatomic potentials ([63, 70] and [74, 75]) were found to melt in the course of the simulations. Further investigation is needed to determine the cause of the low temperature melting phenomenon predicted by these force fields. If one looks at the volumes in Figure 3.5.5 at 300 K, there are several bands of volumes. Close examination of Figure 3.5.1(a) and Figure 3.5.5 reveals that the groupings of average atomic volumes, determined from overall simulation size fluctuations, map directly to the groupings determined from the n-point statistics and PCA analysis. Similar clustering is evident at 900 K, where there is a greater spread in average volumes for the simulations conducted with the different interatomic potentials. The fact that the PC scores automatically capture this effect, without *a* priori information about the phases, bodes well for their utility in capturing high values structure-property linkages. While a simple measure such as the atomic volume would also capture a similar effect, there is no guarantee that it captures all of the significant differences seen in the predicted MD structures. The protocols presented here ensure that all of the salient differences in the ensemble of predicted structures are indeed captured to a high degree of completeness (note that the two PCs referenced in Figure 3.5.4 capture 96.3 % of the differences in the elements of the ensemble).

A.6 Conclusions

This initial study demonstrates the utility and the viability of utilizing rigorous structure quantification protocols to results predicted by MD. Of particular significance is

the fact that similar protocols were previously applied successfully to material structure datasets at the mesoscale. This study reinforces the possibility that a consistent set of structure quantification tools can be designed and applied to a broad range of materials systems at vastly different length/structure scales, and paves the way forward for the formulation and validation of such a universal framework. Furthermore, since the framework employs data-driven approaches, it leads to rigorous, practically useful, low dimensional, representations and visualizations. These are central to our goals for creating high value materials knowledge systems.

A.7 Acknowledgements

SRK and JAG acknowledge support from NIST 70NANB14H191. ZTT and CAB acknowledge support from the NIST Materials Genome Initiative program. The authors would also like to thank Dr. Eric A. Lass for the use of his experimentally-based assessment of Al molar volumes as reference.

No approval or endorsement of any commercial product by NIST is intended or implied. Certain commercial software systems are identified in this paper to facilitate understanding. Such identification does not imply that these software systems are necessarily the best available for the purpose.

APPENDIX B. EXTRACTING KNOWLEDGE FROM MOLECULAR MECHANICS SIMULATIONS OF GRAIN BOUNDARIES USING MACHINE LEARNING

Joshua A. Gomberg [1], Andrew J. Medford [2,3], and Surya R. Kalidindi* [1,3]

[1] School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA

[2] School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA, USA

[3] George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

B.1 Abstract

In this paper, we demonstrate that the "process-structure-property" (PSP) paradigm of materials science can be extended to atomistic grain boundary (GB) simulations through the development of a novel framework that addresses the objective identification of the atoms in the grain boundary regions using the centro-symmetry parameter and local regression, and the quantification of the resulting structure by a pair correlation function (PCF) derived from kernel density estimation (KDE). For asymmetric tilt GBs (ATGBs) in aluminum, models were successfully established connecting the GB macro degrees of freedom (treated as process parameters) and energy (treated as property) to a low-rank GB atomic structure approximation derived from principal component analysis (PCA) of the full ensemble of PCFs aggregated for this study. More specifically, it has been shown that the models produced in this study resulted in average prediction errors less than 13 mJ/m^2, which is less than the error associated with the underlying

simulations when compared with experiments. This demonstration raises the potential for the development and application of PSP linkages from atomistic simulation datasets, and offers a powerful route for extracting high value actionable and transferrable knowledge from such computations.

KEYWORDS: grain boundaries; materials informatics; molecular dynamics; pair correlation function; principal component analysis; process-structure-property linkage

B.2 Introduction

One of the fundamental challenges in materials science is the establishment of high value correlations between the process parameters of a given material and its associated performance characteristics, while accounting for the hierarchical nature of the material's internal structure[29]. Such correlations form the foundations of the field of materials science and engineering, and are generally referred as PSP (process-structure-property) linkages. These linkages are central to all efforts aimed at the development and deployment of new or improved materials for advanced technologies[5, 29, 156-160].

The main hurdle in the establishment of the PSP linkages comes from the fact that the material internal structure spans multiple hierarchical length/structure scales. Furthermore, the rich complexity of features exhibited by different materials at different length/structure scales has greatly impeded the efforts aimed at developing a universally applicable framework. Furthermore, in spite of the tremendous advances made in the experimental characterization of the material internal structure[161, 162], currently available techniques are not yet capable of producing sufficiently large ensembles of

experimentally measured datasets that can be mined for PSP linkages. For this and many other reasons, multiscale models[163-165] offer the most practical path forward for establishing and demonstrating the critical methodologies needed for extracting and validating high value PSP linkages spanning the multiple length/structure scales involved, i.e., the atomic scale to the continuum.

Emerging toolsets of data science and informatics offer tremendous potential for mining the high value PSP linkages from aggregated and curated materials datasets[166-169]. A large fraction of such effort in current literature has only considered relatively simple definitions of the material that included mainly the overall chemical composition of the material. In recent work[28, 30, 170, 171], our research group has championed a new materials data science framework that explicitly accounts for the complex hierarchical material structure. Called Materials Knowledge Systems (MKS)[126, 172-174], this new framework employs spatial correlations to quantify the material structure (at each structure/length scale of interest) and principal component analyses (PCA) to obtain the salient low-dimensional measures needed to represent the complex material structure in the PSP linkages of interest. This new framework has been successfully demonstrated with several case studies dealing with the mesoscale structure of the material [27, 109, 175-178]. Only recently, the application of this framework is being extended to the atomic structure of materials [28, 151].

There is tremendous value in casting the rich, physics-driven, results of the molecular mechanics (MM) or molecular dynamics (MD) simulations as PSP linkages. However, it is not immediately obvious what variables should be selected to describe the process parameters in such linkages. We argue that the process variables selected should

describe the conditions imposed to control or modify the material structure. These might include the thermodynamic ensemble, force fields, and applied loads. At the atomic scale, these can also be captured effectively by the configurational constraints imposed on the material structure. More specifically, the macro degrees of freedom imposed as input in the GB simulations would constitute the process variables. The “structure” would correspond to the elements, configuration and bonding structure of the atoms in a given composition. For atomistic simulations, a commonly employed metric for quantifying structure is the pair correlation function (PCF). The application of MKS framework relies on discretized representation of the material structure, both for quantification of the statistics (i.e., spatial correlations) as well as obtaining low dimensional representations (i.e., PCA). In prior work using microscopy images (e.g., optical, SEM)[31], discrete representations were obtained easily because the image itself is often stored as pixelated values. For point-cloud data such as the results of MD simulations studied here, we need to pay careful attention to how this is accomplished. If the PCFs are computed using the atomic positions directly, they would exhibit very sharp peaks (since the PCF is essentially a weighted sum of Dirac-delta functions located at the specific distances realized in the given atomic structure). This poses two main challenges: (i) The discrete representations of the PCF become very sensitive to the binning, especially as the bin size decreases (in efforts to capture the PCF accurately in their discretized representations). (ii) The discretized representation of the PCF would exhibit a large number of zero values for many of the bins (because of the Dirac-delta nature of the PCFs). Furthermore, if the PCF value for a bin is zero for all the atomic structures studied (this is very likely to happen with any point-cloud datasets), then the PCA can be hindered by rank deficiency

because there is simply no information on that specific bin to compute the corresponding component in all of the orthonormal eigen vectors comprising the PCA basis. Therefore, it is clear that some form of smoothing is essential for the application of the MKS framework on the point-cloud atomic structure datasets. In the present work, this was accomplished using Epanechnikov kernels, which effectively amounts to placing a sphere around each atomic position and then discretizing the volumetric space to obtain discretized, but robust, representations of the PCF useful to establishing the desired PSP linkages.

The concept of a "process-structure" relationship for these atomistic simulations would establish a quantitative connection between the process inputs of the simulation and the resulting atomic-scale structure (output). Previous work has established that molecular force-fields can be classified by the resulting atomic structure using 2-point statistics[28], but there has not been a systematic data-driven effort focused on the extraction of reduced-order "process-structure" linkages capable of rapidly predicting atomic structures as a function of simulation inputs. If such functions can be established in forms that require exceptionally low computational cost for their usage, they offer a unique practical approach for addressing inverse problems where one seeks to identify the process recipes that are likely to result in a desired atomic structure.

Another important type of knowledge produced from molecular dynamics/mechanics simulations can be captured effectively in linkages between atomic-scale structure and a relevant property such as the overall system energy; these linkages may be categorized as "structure-property" relationships. In particular, GB energies play a vital role in the multiscale modeling of materials phenomena, as they serve as a key

input to simulations at a larger scale (e.g., plasticity[34], failure[37], recrystallization[179]). While force-field based calculations are significantly less computationally expensive than their quantum-mechanical counterparts, the datasets often investigated are large in size (10^3 – 10^9 atoms) and high-dimensional, and thus cumbersome for use in multiscale models[34, 37]. Some progress has been made in training machine-learning force fields to results of quantum mechanical methods such as density functional theory for use in molecular dynamics simulations[38, 39, 180, 181], but these methods typically require large training sets (10^3-10^4 systems), and ultimately MD simulations are still necessary to extract knowledge regarding a system. Data-science techniques have also been previously applied for the systematic analysis and knowledge extraction from large MM/MD datasets[41-45] with a focus primarily on proteins and other large biomolecules. Within the materials science community, there has been relatively little effort devoted to a systematic analysis and dimensional reduction of the results of force-field based simulations. This is of particular importance given the recent rise in multiscale and hierarchical methods[29, 46].

It is emphasized here that one of the main benefits of the data science approaches explored in this work is that they facilitate a systematic and effective learning of the deeply embedded knowledge in the numerical datasets produced by MM/MD simulations. In other words, while MM/MD computations are commonly employed to account for the atomic-scale degrees of freedom within a GB structure[40, 182], there is no systematic, data-driven, formalism to capture the knowledge gained from these simulations in forms that allow easy application of the knowledge to new problems. Given the unimaginably large materials space (including all material chemistries and

process variables) that could be covered by the multitude ongoing disparate efforts of researchers everywhere, it behooves us to consider formalisms that allow extraction, fusion, and curation of the knowledge gained from such efforts. Indeed, such an advance is essential to enhance and interpret experimental data, pass information between computational models, and rapidly explore large design spaces (by facilitating solutions to inverse problems of interest). The ability to navigate the potential diversity of GB structures in a low-dimensional space would provide a facile route to rapidly identify structural regions of interest for additional molecular simulations and connect information between the atomic scale simulations and models at larger length scales. Furthermore, recent developments in microscopy have led to the ability to probe directly the atomic scale structures[183], and diffraction techniques that can be used to measure the PCFs;[184, 185] rapid estimation of the energy of an arbitrary GB or atomic structure will allow on-the-fly analysis of these experimental results, providing valuable real-time feedback to the equipment operator[186, 187]. This work aims to establish a foundational data science framework that will facilitate these types of future explorations.

B.3 Dataset

The dataset used in this study was produced by Tschopp et. al [40] and disseminated in an open repository hosted by the NIST Computational File Repository[80]. The use of a publicly available dataset such as this allows multiple research groups (including ours) to apply different techniques and strategies, and to objectively compare the models produced. The reader is referred to previously published papers in literature[81] for details of how this dataset was generated. Here, we summarize only the main details relevant to our study.

The dataset available for the study contained a total of 106 MM simulations that included ATGBs with Σ values of 3, 5, 9, 11 and 13 in aluminum (see Table 4.2.1 – Details of grain boundary simulations used in this study[40, 80].), which reflect the level of coincidence of the atomic structure at the grain boundary. For example, a Σ of 3 indicates that 1/3 of the lattice sites of the two grain orientations that meet at the GB are coincident. All of the simulations employed periodic boundary conditions, with GB planes perpendicular to the Y direction[40].

B.4 Approach for Establishing PSP Linkages at the Atomic Scale

The workflow developed and employed in this work for establishing the structure-property and process-structure linkages of interest is outlined in Figure 4.1.1. Broadly, this workflow depicts three main components: (i) low-rank quantification of the grain boundary atomic structure, (ii) extraction of a structure-property linkage, and (ii) extraction of process-structure linkages. These components are discussed below sequentially, explaining the rationale behind each step involved in each of these components.

B.4.1 Quantification of the Atomic Structure in the GB

The first step in the quantification of the GB atomic structure is the objective identification of the atoms belonging to the GB region in each simulation set. Of the total number of atoms within a given GB simulation (~10^2 to10^5), only a fraction of the total atoms lie in the GB, while the remainder depict the crystalline structure of the bulk. Since only the GB atoms contribute to the GB energy, it is imperative to develop an objective and automated protocol for their identification. Although it might seem that this should

be easy, a set of protocols for differentiating GB atoms from bulk atoms is non-trivial due to the large variety of GB structures and relatively smooth transition between GB and bulk regions.

In this work, the regions associated with the GBs were identified for each simulation using a method based on a local quadratic regression model[52] for the centro-symmetry parameter[51]. This approach relies on centro-symmetry parameter serving as a good surrogate measure of the distortions in the GB atomic structure, and looks at the variation of this scalar parameter as a function of the distance from the GB plane using a local quadratic regression model. The predictor terms used in this regression were the displacements in periodic space (from a given *y*-plane to each atom) raised to both the 1st and 2nd power, as well as an intercept. The regression weights were taken as the probability densities of these displacements being drawn from a normal distribution (equivalent to a Gaussian kernel) with a standard deviation of 4.05 Å. The response variable for this regression model was the square root of the centro-symmetry (CS) parameter of each atom, calculated in a manner consistent with LAMMPS[188]. Taking the square root ensures that both the predictor (displacement) and the response were in units of distance. This can be modeled by the following weighted regression function for each simulation *i*:

$$\sqrt{\omega_{i,a}(y)c_{i,a}} = \sqrt{\omega_{i,a}(y)}\left(\hat{\beta}_{0\mathrm{i}} + \hat{\beta}_{1\mathrm{i}}\delta(y) + \hat{\beta}_{2\mathrm{i}}\delta_{i,a}(y)^2\right) + (\text{error terms}) \qquad (153)$$

where $c_{i,a}$ is the centro-symmetry parameter of atom a, $\omega_{i,a}(y)$ is the regression weight as a function of position, and $\delta_{i,a}(y)$ is the displacement function in periodic space. In the

notation used here, the bolded text used for the regression coefficients refers to the fact that the regression coefficients can be stored in a matrix.

The GB/bulk interfaces were defined to be the local maxima of the 2nd derivative of the modeling equation (the polynomial in parentheses in the right-hand size of Equation 153), or twice the regression coefficient $\widehat{\boldsymbol{\beta}}_{2i}$. An atom was said to be in the GB if it lies within the boundaries defined by the GB/bulk interfaces corresponding to the closest GB. This approach provided computationally fast, objective, and well-defined GB interfaces that aligned well with intuition (see Figure 4.3.1(a)).

The GB atomic structures can be complex and varied, with hundreds or thousands of atoms in the GB per simulation. These structures can be quantified with PCFs[184, 185] or more rigorously with 2-point statistics[28], which are equivalent to directionally resolved PCFs. For atomistic GB simulations, the PCF is a good candidate for use as a structure metric as it is invariant to relative crystal orientation with respect to the reference frame. Only the atoms identified as GB atoms were included in the PCF computation. As PCFs calculated with a traditional binning technique proved too rough (sharp) for model fitting, we employed here a smoothing technique based on kernel density estimation (KDE), as explained next.

For each GB atom, the distances to the 134 nearest neighboring atoms (GB or bulk) were found using the *k*-Nearest Neighbors algorithm[52], and a probability distribution function (PDF) of all neighbor distances for all GB atoms was estimated using KDE[53] with an Epanechnikov kernel. The use of a kernel introduces a smoothing parameter into the PCF; to within a small approximation, a PCF calculated with the

Epanechnikov kernel is equivalent to treating each atom as a uniformly dense sphere of finite radius. The Epanechnikov kernel for bandwidth *h* and distance *u* along the PDF is:

$$\kappa^{\mathrm{e}}(u,h) = \begin{cases} \dfrac{3}{4\sqrt{5h^2}}\left(1 - \dfrac{u^2}{5h^2}\right) & \text{for } (u/h)^2 < 5 \\ 0 & \text{otherwise} \end{cases} \tag{154}$$

This kernel can be used to construct a PDF using the following equation:

$$\psi_i(r) = \frac{1}{134 \times N_i^G} \sum_{a \in \mathcal{Z}_i^G} \sum_{k=1}^{134} \kappa^{\mathrm{e}}\left(r - \left\|\vec{R}_{i,a}^{(k)}\right\|, h_{\mathrm{e}}\right) \tag{155}$$

In this equation, $\mathcal{Z}_i^{\mathrm{G}}$ represents the set of atoms in the GB, N_i^{G} represents the number of atoms in this set, h_{e} is the Epanechnikov kernel bandwidth, and $\left\|\vec{R}_{i,a}^{(k)}\right\|$ is the magnitude of the displacement vector from atom *a* to its *k*th nearest neighbor. A PCF can be expressed in terms of this PDF scaled by the inverse squared distance and appropriate constants. The formulation of the PCF used in this study can be expressed as:

$$\gamma_i(r) = \frac{134}{4\pi r^2 n_0}\psi_i(r) \tag{156}$$

where, n_0 is the atomic number density of bulk crystalline Al (6.02×10^{-2} Å^{-3}). For each simulation, the value of the PCF was calculated for 512 equally spaced points from 0 to 6.3 Å. The cutoff radius of 6.3 Å was chosen as it corresponds to the cutoff distance associated with the interatomic potential used in these simulations.[189] Figure 4.3.1(b) depicts an example of one such PCF, in this case a Σ9 simulation with an inclination

angle (θ) of 22.99°. For ATGBs, the inclination angle is the angle between the GB plane and the plane of reflection symmetry between the two crystal lattices.

Although the PCF represents a detailed quantification of the atomic structure, this is still a high-dimensional data structure (equal to the number of points where the PCF is sampled) to allow computationally efficient comparison of different GBs and the establishment of correlations with either the GB energy or the GB macro degrees of freedom. This is where dimensionality reduction techniques can prove valuable. After subtracting the mean PCF from the discretely sampled PCFs of all simulations, principal component analysis (PCA) [49] was performed via the singular value decomposition. PCA is a common technique for dimensional reduction that determines an orthogonal basis for the data where i^{th} eigenvector corresponds to the direction with the i^{th} largest variance, as illustrated in Figure 4.5.2(a).

In this case, the input dataset for PCA is the entire ensemble of PCFs for all Al ATGB simulations, and Figure 4.5.2(b) depicts eigenvectors 1,2,3, and 6 of the dataset. Truncating a PCA representation of a structure at the k^{th} value/vector yields the best possible rank-k approximation to the full dataset. This provides a systematic way to represent a high-dimensional structure in a low-dimensional space while still preserving a well-defined amount of variance from the entire system. After a low-dimensional subspace has been defined, all datasets can naturally be projected onto this subspace; the new coordinates provide their PC scores. Furthermore, given any arbitrary point in this PC space it is straightforward to reconstruct the PCF corresponding to the point by using a linear combination of the appropriate PC vectors weighted with the scores corresponding to the point in low-dimensional space. Thus, the low-dimensional

representation of the atomic structure provides a route for not only analyzing existing datasets, but also for predicting full atomic structures with properties interpolated between the given data using, e.g., reverse Monte Carlo methods.

B.4.2 Structure-Property Linkages

In order to establish a quantitative structure-property relationship in this work, multivariate linear regression was applied. While the PCs are ranked in descending order by the amount of retained variance in the PCF, this is not necessarily the same order for best explaining the variance of GB energy in a predictive regression model. In this model, the GB energy associated with the structure serves as the response variable. First, separate 1-PC regression models for predicting GB energy from each PC individually were constructed. Since the PCs are orthogonal, the best 2-PC regression model in terms of mean squared error is the sum of the two 1-PC models with the smallest individual mean squared errors; as such, the regression coefficients for each PC in the 2-component model are identical to the regression coefficients for each of the corresponding 1-component models. For this data set, the predictors of best two-component structure-property regression model for predicting GB energy are the scores associated with PCs 1 and 6 (see Figure 4.5.2(c)).

In order to verify the robustness of the model, 3-fold cross validation was applied. In 3-fold cross validation, the data set is first randomly divided into three roughly equal groups. Next, three separate models are fitted to the data within every possible combination of two groups. The absolute prediction error is calculated for each model using the data from the group excluded from model fitting. The mean absolute cross-

validation error is the sum of the total prediction error from the three models divided by the total number of data points. This cross-validation procedure also ensures that overfitting of the model is not an issue.

B.4.3 Process-Structure Linkages

For the purpose of establishing process-structure linkages, the "process" of simulation is represented by θ, which corresponds to one of the five macro degrees of freedom, and Σ value, which constrains another three degrees of freedom.[190] As the twist angle is fixed at zero for ATGBs, the resulting model accounts for all five degrees of freedom. To establish the process-structure relationship, for every set of simulations with the same Σ value, separate 3rd order polynomial regression models were constructed that use GB inclination angle (θ) to predict the PC scores. The use of separate models for each Σ value can be justified because for this data set, each Σ is represented by a single coincidence site lattice (CSL). For data sets where this is not the case, separate models should be constructed for each unique CSL.

B.5 Results and Discussion

The Epanechnikov kernel bandwidth serves as a modeling hyperparameter, and the value chosen for the PCF calculation (0.42 Å) corresponds to the error minimum in the full PSP model and full dataset, as illustrated in Figure 4.5.1. This is roughly equivalent to treating each atom as a uniformly dense sphere with a radius of 0.42*√5=0.94Å (about 2/3 the atomic radius of crystalline Al) rather than as a point particle. However, given the relatively wide well depicted in Figure 4.5.1, any bandwidth from ca. 0.35 – 0.5 would

increase the error by < 1 mJ/m^2, indicating that overfitting is not a concern for this hyperparameter selection.

We first examine the structure-property model obtained by regression between the PC values and the associated grain boundary energy. The R^2 value associated with our two-component (PCs 1 and 6) structure-property regression model is 0.98, with PC 1 accounting for 96% of the explained variance. The inclusion of the third-best principal component in terms of fitting (PC 4) did not appreciably improve the fit. The fitting function is presented in Equation 157, where $\hat{E}_i$ is the predicted energy and $\mathbf{T}_{i1}$ and $\mathbf{T}_{i6}$ are the scores corresponding to PCs 1 and 6, respectively.

$$\hat{\mathrm{E}}_{\mathrm{i}}\ (\mathrm{mJ\ m^{-2}}) \approx -548.66 \cdot \mathrm{T}_{\mathrm{i1}} - 1070.89 \cdot \mathrm{T}_{\mathrm{ij}} + 387.55 \tag{157}$$

Figure 4.6.1(a) shows a parity plot of the GB energies predicted by the regression model plotted against the values computed from the full simulations. The mean absolute value of the error in prediction is roughly 11.4 mJ/m^2. This is substantially less than the error in prediction of simulation versus experiment over a large range of θ as shown in the original dataset (see Ref. [40], Fig 1).

Figure 4.6.1(b) shows a Box-Whisker plot of the mean absolute errors from 1000 instances of 3-fold cross-validation. The box represents the interquartile range, and the dashed ‘whiskers’ have a length 1.5 times that of the interquartile range; points outside this range represented as dots are considered outliers. This shows that, even in the case of extreme outliers, the prediction error is still fairly small (<13 mJ/m^2). Given that the accuracy of GB energies computed with force-field models can be on the order of 50

mJ/m^2 when compared to experiment,[40] this result indicates that the method will be able to serve as an efficient and reliable surrogate to expensive molecular mechanics models for GB energies. Additionally, this plot confirms that the model has not been over-fitted.

The accuracy of the structure-property relationship is remarkable given its simplicity. The PCF represents a substantial compression of information from the full atomic structure, and the fact that the PCF is correlated to the energy through a linear model with only two principal components is noteworthy. However, inspection of the underlying interatomic potential yields some insight into this finding. The potential applied in this work is based on the embedded atom model[191, 192]:

$$E^{\mathrm{tot}} = \frac{1}{2}\sum_{\mathrm{a}}\sum_{\tilde{\mathrm{a}}\neq\mathrm{a}} \Phi_{\mathrm{a}\tilde{\mathrm{a}}}(\mathrm{r}_{\mathrm{a}\tilde{\mathrm{a}}}) + \sum_{\mathrm{a}} \eta_{\mathrm{a}}\left(\sum_{\tilde{\mathrm{a}}\neq\mathrm{a}} \rho_{\tilde{\mathrm{a}}}(\mathrm{r}_{\mathrm{a}\tilde{\mathrm{a}}})\right) \tag{158}$$

where E^{tot} is the energy of the system, $\mathbf{\Phi}_{a\tilde{a}}$ is an interatomic pair potential, $\rho_{\tilde{a}}$ is the "atomic electron density" function, η_a is the embedding energy function, and $\mathbf{r}_{a\tilde{a}}$ is the distance between atoms a and $\tilde{a}$. Inspection of the expression reveals that the interatomic distance, $\mathbf{r}_{a\tilde{a}}$, is the fundamental variable of both summation terms. The fact that the PCF is a function of the probability distribution of $\mathbf{r}_{a\tilde{a}}$ (see Equation 156) provides a strong mathematical justification for its ability to accurately predict energies based on the embedded atom model, and suggests that the accuracy of the approach will generalize to other similar potentials where the interatomic distance is the fundamental variable[83, 84]. However, it should be noted that a regression model constructed using PCFs calculated by a traditional binning technique had relatively weak predictive power. This means that a predictive model must be robust against structural variance as pertaining to

small local changes in atomic position, which are diminished by the smoothing parameter in KDE-derived PCFs.

Next, we analyze the model used to create a process-structure linkage. Employing the terms of a 3rd order polynomial of θ as predictors, separate regression models were constructed for each Σ value to estimate the scores of PC 1 and 6 (see Table 4.6.1).

Figure 4.6.2 shows the PC 1 and 6 scores along with the scores predicted from the regression models. For both PC 1 and 6, the regression fit for Σ = 5 and 13 are much poorer than the fit for Σ = 3, 9 and 11. ATGBs with Σ = 5 and 13 and Σ = 3, 9 and 11 have misorientation axes of [001] and [110], respectively. The misorientation axis is the axis about which the lattices on either side of the GB are rotated to bring them into coincidence. This suggests that different misorientation axes influence the orientation of the PC vectors in a complex manner that is not fully described in this simple model. However, this added complexity does not manifest itself in the previously described structure-energy relation.

The method outlined here provides a framework for efficiently extracting quantitative and transferable PSP linkages from molecular mechanics/dynamics simulations. Figure 4.6.3 illustrates the continuous nature of these linkages. The structure-property relationship illustrated in Figure 4.6.3(a) can predict the GB energy for any Al ATGB with a reasonably similar structure to those in the model. The process-structure relationship can predict the structure itself as a function of θ and Σ (see Figure 4.6.3(b)). These linkages will aid in the coupling of complex GB boundary structures into multiscale models where hundreds or thousands of different GB structures may arise.

The applicability of models such as the ones discussed here is not limited to force-field based simulations. Both force-field based and *ab initio* atomic simulation techniques are designed to model a highly nonlinear landscape of potential atomic structures. However, if a data set is limited to a sufficiently restricted set of potential atomic structures, linear regression models should be able to accurately explain their properties, although we note that properties with strong directionality (such as covalent bonding) will likely require descriptors which can directionally resolve the atomic environments[28, 151, 181]. The PCF's used here are spherically symmetric and are hence expected to perform best for the relatively homogenous bonding of metals.

B.6 Conclusions

Here, data-driven learning models such as those employed in MKS have been successfully adapted to MD simulations of aluminum ATGBs. Quantitative linkages such as those established in this work present opportunities for advanced GB engineering, faster global optimization of GB structures, and real-time integration of computational and experimental results. Future work will focus on establishing the generality of the technique for GBs in other common structural materials (e.g., Cu, Fe), exploring the possibility of analyzing more chemically heterogeneous systems (e.g., alloys, oxides), and quantification of uncertainty. Given the success of the current model which includes only two principal components and a linear regression model, it is anticipated that the approach will be widely applicable.

B.7 Acknowledgements

Work was supported by the National Institute for Standards and Technology (No. 70NANB14H191).

REFERENCES

[1] S. Ghosh, K. Lee, S. Moorthy, Multiple scale analysis of heterogeneous elastic structures using homogenization theory and voronoi cell finite element method, International Journal of Solids and Structures 32(1) (1995) 27-62.

[2] V.G. Kouznetsova, M.G.D. Geers, W.A.M. Brekelmans, Multi-scale second-order computational homogenization of multi-phase materials: a nested finite element solution strategy, Computer Methods in Applied Mechanics and Engineering 193(48–51) (2004) 5525-5550.

[3] H. Kadowaki, W.K. Liu, Bridging multi-scale method for localization problems, Computer Methods in Applied Mechanics and Engineering 193(30–32) (2004) 3267-3302.

[4] D.J. Luscher, D.L. McDowell, C.A. Bronkhorst, A second gradient theoretical framework for hierarchical multiscale modeling of materials, International Journal of Plasticity 26(8) (2010) 1248-1275.

[5] B.L. Adams, S.R. Kalidindi, D.T. Fullwood, Microstructure sensitive design for performance optimization, Oxford, Elsevier Science, 2012.

[6] S.R. Kalidindi, Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials, International Materials Reviews 60(3) (2015) 150-168.

[7] K. Kirane, S. Ghosh, M. Groeber, A. Bhattacharjee, Grain Level Dwell Fatigue Crack Nucleation Model for Ti Alloys Using Crystal Plasticity Finite Element Analysis, Journal of Engineering Materials and Technology-Transactions of the Asme 131(2) (2009).

[8] C.P. Przybyla, D.L. McDowell, Simulated Microstructure-Sensitive Extreme Value Probabilities for High Cycle Fatigue of Duplex Ti-6Al-4V, International Journal of Plasticity, Special Issue in Honor or Nobutada Ohno (2012).

[9] B.L. Wang, Y.H. Wen, J. Simmons, Y.Z. Wang, Systematic approach to microstructure design of Ni-base alloys using classical nucleation and growth relations coupled with phase field modeling, Metallurgical and Materials Transactions a-Physical Metallurgy and Materials Science 39A(5) (2008) 984-993.

[10] Y.H. Wen, J.P. Simmons, C. Shen, C. Woodward, Y. Wang, Phase-field modeling of bimodal particle size distributions during continuous cooling, Acta Materialia 51(4) (2003) 1123-1132.

[11] S.R. Kalidindi, A. Bhattacharya, R. Doherty, Detailed Analysis of Plastic Deformation in Columnar Polycrystalline Aluminum Using Orientation Image Mapping and Crystal Plasticity Models, Proceedings of the Royal Society of London: Mathematical, Physical and Engineering Sciences. 460(2047) (2004) 1935 - 1956

[12] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, G. Ceder, Predicting Crystal Structures with Data Mining of Quantum Calculations, Physical Review Letters 91(13) (2003) 135503.

[13] C.A. Becker, J. Ågren, M. Baricco, Q. Chen, S.A. Decterov, U.R. Kattner, J.H. Perepezko, G.R. Pottlacher, M. Selleby, Thermodynamic modelling of liquids: CALPHAD approaches and contributions from statistical physics, physica status solidi (b) 251(1) (2014) 33-52.

[14] Z.T. Trautt, Y. Mishin, Capillary-driven grain boundary motion and grain rotation in a tricrystal: A molecular dynamics study, Acta Materialia 65 (2014) 19-31.

[15] A. Karma, Z.T. Trautt, Y. Mishin, Relationship between Equilibrium Fluctuations and Shear-Coupled Motion of Grain Boundaries, Physical Review Letters 109(9) (2012).

[16] Z.T. Trautt, M. Upmanyu, A. Karma, Interface mobility from interface random walk, Science 314(5799) (2006) 632-635.

[17] M. Palumbo, B. Burton, A.C.E. Silva, B. Fultz, B. Grabowski, G. Grimvall, B. Hallstedt, O. Hellman, B. Lindahl, A. Schneider, P.E.A. Turchi, W. Xiong, Thermodynamic modelling of crystalline unary phases, Phys Status Solidi B 251(1) (2014) 14-32.

[18] P. Derosa, T. Cagin, Multiscale Modeling: From Atoms to Devices, CRC Press2010.

[19] D. Sholl, J.A. Steckel, Density functional theory: a practical introduction, John Wiley & Sons2011.

[20] D.C. Rapaport, The art of molecular dynamics simulation, Cambridge university press2004.

[21] E. Moeendarbary, T.Y. Ng, M. Zangeneh, Dissipative Particle Dynamics: Introduction, Methodology and Complex Fluid Applications - A Review, International Journal of Applied Mechanics 01(04) (2009) 737-763.

[22] O.C. Zienkiewicz, R.L. Taylor, O.C. Zienkiewicz, R.L. Taylor, The finite element method, McGraw-hill London1977.

[23] W.S. Cleveland, Data science: an action plan for expanding the technical areas of the field of statistics, International statistical review 69(1) (2001) 21-26.

[24] V. Dhar, Data science and prediction, Communications of the ACM 56(12) (2013) 64-73.

[25] T. Hey, S. Tansley, K. Tolle, The fourth paradigm, Data-Intensive Scientific Discovery. Microsoft Research (2009).

[26] C. Anderson, The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine, Updated 6/23/2008), Available at: http://www. wired. com/science/discoveries/magazine/16-07/pb_theory, 2008.

[27] Y.C. Yabansu, P. Steinmetz, J. Hötzer, S.R. Kalidindi, B. Nestler, Extraction of reduced-order process-structure linkages from phase-field simulations, Acta Materialia 124 (2017) 182-194.

[28] S.R. Kalidindi, J.A. Gomberg, Z.T. Trautt, C.A. Becker, Application of data science tools to quantify and distinguish between structures and models in molecular dynamics datasets, Nanotechnology 26(34) (2015) 344006.

[29] S. Kalidindi, Hierarchical materials informatics : novel analytics for materials data, Butterworth-Heinemann, Waltham, MA 2015.

[30] A. Gupta, A. Cecen, S. Goyal, A.K. Singh, S.R. Kalidindi, Structure–property linkages using a data science approach: Application to a non-metallic inclusion/steel composite system, Acta Materialia 91 (2015) 239-254.

[31] S.R. Kalidindi, Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials, International Materials Reviews 60(3) (2015) 150-168.

[32] A. Çeçen, T. Fast, E.C. Kumbur, S.R. Kalidindi, A data-driven approach to establishing microstructure–property relationships in porous transport layers of polymer electrolyte fuel cells, Journal of Power Sources 245(0) (2014) 144-153.

[33] A. Çeçen, E.A. Wargo, A.C. Hanna, D.M. Turner, S.R. Kalidindi, E.C. Kumbur, 3-D Microstructure Analysis of Fuel Cell Materials: Spatial Distributions of Tortuosity, Void Size and Diffusivity, Journal of The Electrochemical Society 159(3) (2012) B299-B307.

[34] D.L. McDowell, A perspective on trends in multiscale plasticity, International Journal of Plasticity 26(9) (2010) 1280-1309.

[35] R. Phillips, M. Dittrich, K. Schulten, Quasicontinuum Representations of Atomic-Scale Mechanics: From Proteins to Dislocations, Annual Review of Materials Research 32(1) (2002) 219-233.

[36] D.L. Olmsted, E.A. Holm, S.M. Foiles, Survey of computed grain boundary properties in face-centered cubic metals-II: Grain boundary mobility, Acta Materialia 57(13) (2009) 3704-3713.

[37] C.L. Rountree, R.K. Kalia, E. Lidorikis, A. Nakano, L. Van Brutzel, P. Vashishta, Atomistic Aspects of Crack Propagation in Brittle Materials: Multimillion Atom Molecular Dynamics Simulations, Annual Review of Materials Research 32(1) (2002) 377-400.

[38] J. Behler, M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Physical Review Letters 98(14) (2007) 146401.

[39] M. Rupp, M.R. Bauer, R. Wilcken, A. Lange, M. Reutlinger, F.M. Boeckler, G. Schneider, Machine learning estimates of natural product conformational energies, PLoS Comput Biol 10(1) (2014) e1003400.

[40] M.A. Tschopp, S.P. Coleman, D.L. McDowell, Symmetric and asymmetric tilt grain boundary structure and energy in Cu and Al (and transferability to other fcc metals), Integrating Materials and Manufacturing Innovation 4(1) (2015) 1-14.

[41] A.E. García, Large-amplitude nonlinear motions in proteins, Physical review letters 68(17) (1992) 2696.

[42] A. Amadei, A.B.M. Linssen, H.J.C. Berendsen, Essential dynamics of proteins, Proteins: Structure, Function, and Bioinformatics 17(4) (1993) 412-425.

[43] R. Hegger, A. Altis, P.H. Nguyen, G. Stock, How Complex Is the Dynamics of Peptide Folding?, Physical Review Letters 98(2) (2007) 028102.

[44] P.I. Zhuravlev, C.K. Materese, G.A. Papoian, Deconstructing the Native State: Energy Landscapes, Function, and Dynamics of Globular Proteins, The Journal of Physical Chemistry B 113(26) (2009) 8800-8812.

[45] P. Das, M. Moll, H. Stamati, L.E. Kavraki, C. Clementi, Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction, Proceedings of the National Academy of Sciences 103(26) (2006) 9885-9890.

[46] D.L. McDowell, G.B. Olson, Concurrent design of hierarchical materials and structures, Scientific Modeling and Simulation SMNS 15(1) (2008) 207-240.

[47] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.-J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21(13) (2005) 3043-3048.

[48] J.J. Jansen, H.C.J. Hoefsloot, J. van der Greef, M.E. Timmerman, J.A. Westerhuis, A.K. Smilde, ASCA: analysis of multivariate data obtained from an experimental design, Journal of Chemometrics 19(9) (2005) 469-481.

[49] I.T. Jolliffe, Principal component analysis, 2nd ed.. ed., New York : Springer-Verlag, New York, 2002.

[50] H.J. Nussbaumer, Fast Fourier Transform and Convolution Algorithms, Springer Berlin Heidelberg2013.

[51] C.L. Kelchner, S.J. Plimpton, J.C. Hamilton, Dislocation nucleation and defect structure during surface indentation, Physical Review B 58(17) (1998) 11085-11088.

[52] T. Hastie, The elements of statistical learning data mining, inference, and prediction, 2nd ed.. ed., Springer, New York, 2009.

[53] B.W. Silverman, Density estimation for statistics and data analysis, Chapman and Hall, New York, 1986.

[54] P.d.B. Harrington, N.E. Vieira, J. Espinoza, J.K. Nien, R. Romero, A.L. Yergey, Analysis of variance–principal component analysis: A soft tool for proteomic discovery, Analytica chimica acta 544(1) (2005) 118-127.

[55] P. Vik, Regression, ANOVA, and the general linear model: A statistics primer, SAGE Publications2013.

[56] A. Dean, D. Voss, Design and analysis of experiments, Springer-Verlag NY, Inc1999.

[57] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley2003.

[58] J. Ramsay, B.W. Silverman, Functional Data Analysis, Springer2005.

[59] J.E. Gentle, Matrix Algebra: Theory, Computations, and Applications in Statistics, Springer New York2010.

[60] G.P. Purja Pun, Y. Mishin, to be published, (2013).

[61] M.I. Mendelev, D.J. Srolovitz, G.J. Ackland, S. Han, Effect of Fe Segregation on the Migration of a Non-Symmetric Σ5 Tilt Grain Boundary in Al, Journal of Materials Research 20(01) (2005) 208-218.

[62] M.I. Mendelev, M. Asta, M.J. Rahman, J.J. Hoyt, Development of interatomic potentials appropriate for simulation of solid–liquid interface properties in Al–Mg alloys, Philosophical Magazine 89(34-36) (2009) 3269-3285.

[63] D. Schopf, P. Brommer, B. Frigan, H.-R. Trebin, Embedded atom method potentials for Al-Pd-Mn phases, Physical Review B 85(5) (2012) 054201.

[64] A. Landa, P. Wynblatt, D.J. Siegel, J.B. Adams, O.N. Mryasov, X.Y. Liu, Development of glue-type potentials for the Al–Pb system: phase diagram calculation, Acta Materialia 48(8) (2000) 1753-1761.

[65] X.-Y. Liu, F. Ercolessi, J.B. Adams, Aluminium interatomic potential from density functional theory calculations with improved stacking fault energy, Modelling and Simulation in Materials Science and Engineering 12(4) (2004) 665.

[66] M.I. Mendelev, M.J. Kramer, C.A. Becker, M. Asta, Analysis of semi-empirical interatomic potentials appropriate for simulation of crystalline and liquid Al and Cu, Philosophical Magazine 88(12) (2008) 1723-1750.

[67] Y. Mishin, D. Farkas, M.J. Mehl, D.A. Papaconstantopoulos, Interatomic potentials for monoatomic metals from experimental data and ab initio calculations, Physical Review B 59(5) (1999) 3393-3407.

[68] J.B. Sturgeon, B.B. Laird, Adjusting the melting point of a model system via Gibbs-Duhem integration: Application to a model of aluminum, Physical Review B 62(22) (2000) 14720-14727.

[69] J.M. Winey, K. Alison, Y.M. Gupta, A thermodynamic approach to determine accurate potentials for molecular dynamics simulations: thermoelastic response of aluminum, Modelling and Simulation in Materials Science and Engineering 17(5) (2009) 055004.

[70] X.W. Zhou, R.A. Johnson, H.N.G. Wadley, Misfit-energy-increasing dislocations in vapor-deposited CoFe/NiFe multilayers, Physical Review B 69(14) (2004) 144113.

[71] R.R. Zope, Y. Mishin, Interatomic potentials for atomistic simulations of the Ti-Al system, Physical Review B 68(2) (2003) 024102.

[72] X.-Y. Liu, P.P. Ohotnicky, J.B. Adams, C.L. Rohrer, R.W. Hyland Jr, Anisotropic surface segregation in Al・Mg alloys, Surface Science 373(2–3) (1997) 357-370.

[73] M.P. Allen, D.J. Tildesley, Computer simulation of liquids, Oxford Oxfordshire, Clarendon Press ;, 1987.

[74] J.E. Angelo, N.R. Moody, M.I. Baskes, Trapping of hydrogen to lattice defects in nickel, Modelling and Simulation in Materials Science and Engineering 3(3) (1995) 289.

[75] M.I. Baskes, X. Sha, J.E. Angelo, N.R. Moody, Trapping of hydrogen to lattice defects in nickel, Modelling and Simulation in Materials Science and Engineering 5(6) (1997) 651.

[76] Y. Mishin, M.J. Mehl, D.A. Papaconstantopoulos, Embedded-atom potential for B2-NiAl, Physical Review B 65(22) (2002) 224114.

[77] Y. Mishin, Atomistic modeling of the γ and γ′-phases of the Ni–Al system, Acta Materialia 52(6) (2004) 1451-1467.

[78] G.P. Purja Pun, Y. Mishin, Development of an interatomic potential for the Ni-Al system, Philosophical Magazine 89(34-36) (2009) 3245-3267.

[79] G.P. Purja Pun, Y. Mishin, to be published, (2013).

[80] M. Tschopp, S. Coleman, D.L. McDowell, Al-Cu Symmetric/Asymmetric Tilt Grain Boundary Dataset, http://hdl.handle.net/11256/358.

[81] M.A. Tschopp, D.L. McDowell, Structures and energies of Sigma 3 asymmetric tilt grain boundaries in copper and aluminium, Philos Mag 87 (2007).

[82] R. Szeliski, Computer vision: algorithms and applications, Springer Science & Business Media2010.

[83] M. Finnis, J. Sinclair, A simple empirical N-body potential for transition metals, Philosophical Magazine A 50(1) (1984) 45-55.

[84] F. Ercolessi, M. Parrinello, E. Tosatti, Simulation of gold in the glue model, Philosophical magazine A 58(1) (1988) 213-226.

[85] W.T. Read, W. Shockley, Dislocation Models of Crystal Grain Boundaries, Physical Review 78(3) (1950) 275-289.

[86] L. Priester, Grain boundaries: from theory to engineering, Springer Science & Business Media2012.

[87] H. Samet, Foundations of Multidimensional and Metric Data Structures, Elsevier/Morgan Kaufmann2006.

[88] D. Meagher, Geometric modeling using octree encoding, Computer Graphics and Image Processing 19(2) (1982) 129-147.

[89] J. Leskovec, A. Rajaraman, J.D. Ullman, Mining of Massive Datasets, Cambridge University Press2014.

[90] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, Journal of the American Chemical Society 117(19) (1995) 5179-5197.

[91] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, Journal of computational chemistry 4(2) (1983) 187-217.

[92] N.L. Allinger, Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms, Journal of the American Chemical Society 99(25) (1977) 8127-8134.

[93] Materials Genome Initiative for Global Competitiveness., in: N.S.a.T. Council (Ed.) 2011.

[94] T.M. Pollock, J.E. Allison, D.G. Backman, M.C. Boyce, M. Gersh, E.A. Holm, R. LeSar, M. Long, A.C.P. IV, J.J. Schirra, D.D. Whitis, C. Woodward, Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security, The National Acamedies Press, Washington DC, 2008.

[95] D.L. McDowell, T.L. Story, New Directions in Materials Design Science and Engineering (MDS&E), Report of a NSF DMR-sponsored workshop, 1998, pp. October 19-21.

[96] A National Strategic Plan for Advanced Manufacturing, in: N.S.a.T.C. Executive Office of the President (Ed.) Feb 2012.

[97] G.J. Schmitz, U. Prahl, ICMEg–the Integrated Computational Materials Engineering expert group–a new European coordination action, Integrating Materials and Manufacturing Innovation 3(1) (2014) 2.

[98] G. Linden, B. Smith, J. York, Amazon. com recommendations: Item-to-item collaborative filtering, Internet Computing, IEEE 7(1) (2003) 76-80.

[99] I. Li, A. Dey, J. Forlizzi, A stage-based model of personal informatics systems, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2010, pp. 557-566.

[100] M. Hohman, K. Gregory, K. Chibale, P.J. Smith, S. Ekins, B. Bunin, Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery, Drug discovery today 14(5) (2009) 261-270.

[101] J.M. Tien, Toward a decision informatics paradigm: a real-time, information-based approach to decision making, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 33(1) (2003) 102-113.

[102] T.T. Wan, Healthcare informatics research: from data to evidence-based management, Journal of Medical Systems 30(1) (2006) 3-7.

[103] B.L. Adams, T. Olson, Mesostructure - properties linkage in polycrystals, Progress in Materials Science 43(1) (1998) 1-88.

[104] P.P. Castañeda, J.J. Telega, B. Gambin, Nonlinear homogenization and its applications to composites, polycrystals and smart materials, Springer2004.

[105] D.T. Fullwood, B.L. Adams, S.R. Kalidindi, A strong contrast homogenization formulation for multi-phase anisotropic materials, Journal of the Mechanics and Physics of Solids 56(6) (2008) 2287-2297.

[106] G.W. Milton, The Theory of Composites Cambridge Monographs on Applied and Computational Mathematics (2001).

[107] F. Roters, P. Eisenlohr, L. Hantcherli, D.D. Tjahjanto, T.R. Bieler, D. Raabe, Overview of constitutive laws, kinematics, homogenization and multiscale methods in crystal plasticity finite-element modeling: Theory, experiments, applications, Acta Materialia 58(4) (2010) 1152-1211.

[108] J.R. Willis, Variational and related methods for the overall porperties of composite materials, Advances in Applied Mechanics 21 (1981) 2-78.

[109] A. CeCen, T. Fast, E.C. Kumbur, S.R. Kalidindi, A Data-driven Approach to Establishing Microstructure-Property Relationships in Porous Transport Layers of Polymer Electrolyte Fuel Cells, Journal of Power Sources 245 (2014) 144-153.

[110] S.R. Kalidindi, S.R. Niezgoda, A.A. Salem, Microstructure informatics using higher-order statistics and efficient data-mining protocols, JOM 63(4) (2011) 34-41.

[111] S.R. Kalidindi, S.R. Niezgoda, G. Landi, S. Vachhani, T. Fast, A Novel Framework for Building Materials Knowledge Systems, Computers Materials & Continua 17(2) (2010) 103-125.

[112] G. Landi, S.R. Kalidindi, Thermo-Elastic Localization Relationships for Multi-Phase Composites, Cmc-Computers Materials & Continua 16(3) (2010) 273-293.

[113] G. Landi, S.R. Niezgoda, S.R. Kalidindi, Multi-scale modeling of elastic response of three-dimensional voxel-based microstructure datasets using novel DFT-based knowledge systems, Acta Materialia 58(7) (2010) 2716-2725.

[114] T. Fast, S.R. Niezgoda, S.R. Kalidindi, A new framework for computationally efficient structure-structure evolution linkages to facilitate high-fidelity scale bridging in multi-scale materials models, Acta Materialia 59(2) (2011) 699-707.

[115] T. Fast, S.R. Kalidindi, Formulation and Calibration of Higher-Order Elastic Localization Relationships Using the MKS Approach Acta Materialia 59 (2011) 4595-4605.

[116] S.R. Kalidindi, Computationally-Efficient Fully-Coupled Multi-Scale Modeling of Materials Phenomena Using Calibrated Localization Linkages, ISRN Materials Science (2012).

[117] Y.C. Yabansu, D.K. Patel, S.R. Kalidindi, Calibrated Localization Relationships for Elastic Response of Polycrystalline Aggregates, Acta Materialia 81 (2014) 151-160.

[118] J.B. Shaffer, M. Knezevic, S.R. Kalidindi, Building texture evolution networks for deformation processing of polycrystalline fcc metals using spectral approaches: Applications to process design for targeted performance, International Journal of Plasticity 26(8) (2010) 1183-1194.

[119] H.F. Al-Harbi, S.R. Kalidindi, Crystal plasticity finite element simulations using a database of discrete Fourier transforms, International Journal of Plasticity (2014).

[120] H.F. Al-Harbi, G. Landi, S.R. Kalidindi, Multi-scale modeling of the elastic response of a structural component made from a composite material using the materials knowledge system, Modelling and Simulation in Materials Science and Engineering 20(5) (2012) 055001.

[121] S. Litster, W.K. Epting, E.A. Wargo, S.R. Kalidindi, E.C. Kumbur, Morphological Analyses of Polymer Electrolyte Fuel Cell Electrodes with Nano-Scale Computed Tomography Imaging, Fuel Cells 13(5) (2013) 935-945.

[122] E.A. Wargo, V.P. Schulz, A. Çeçen, S.R. Kalidindi, E.C. Kumbur, Resolving macro- and micro-porous layer interaction in polymer electrolyte fuel cells using focused ion beam and X-ray computed tomography, Electrochimica Acta 87(0) (2013) 201-212.

[123] E.A. Wargo, A.C. Hanna, A. Cecen, S.R. Kalidindi, E.C. Kumbur, Selection of Representative Volume Elements for Pore-Scale Analysis of Transport in Fuel Cell Materials, Journal of Power Sources 197 (2012) 168-179.

[124] B.L. Adams, S.R. Kalidindi, D. Fullwood, Microstructure Sensitive Design for Performance Optimization, Butterworth-Heinemann2012.

[125] D.L. McDowell, J.H. Panchal, H.-J. Choi, C.C. Seepersad, J.K. Allen, F. Mistree, Integrated Design of Multiscale, Multifunctional Materials and Products, Elsevier2009.

[126] J.H. Panchal, S.R. Kalidindi, D.L. McDowell, Key computational modeling issues in integrated computational materials engineering, Computer-Aided Design 45(1) (2013) 4-25.

[127] D.T. Fullwood, S.R. Niezgoda, B.L. Adams, S.R. Kalidindi, Microstructure sensitive design for performance optimization, Progress in Materials Science 55(6) (2010) 477-562.

[128] A.J. Schwartz, M. Kumar, B.L. Adams, Electron Backscatter Diffraction in Materials Science, Kluwer Academic/Plenum Publishers, New York (2000).

[129] S.M. Qidwai, D.M. Turner, S.R. Niezgoda, A.C. Lewis, A.B. Geltmacher, D.J. Rowenhorst, S.R. Kalidindi, Estimating response of polycrystalline materials using sets of weighted statistical volume elements (WSVEs), Acta Materialia 60 (2012) 5284–5299.

[130] D.J. Rowenhorst, A.C. Lewis, G. Spanos, Three-dimensional analysis of grain topology and interface curvature in a b-titanium alloy, Acta Materialia 58(16) (2010) 5511-5519.

[131] S. Torquato, Random Heterogeneous Materials, Springer-Verlag, New York, 2002.

[132] W.F. Brown, Solid Mixture Permittivities, The Journal of Chemical Physics 23(8) (1955) 1514-1517.

[133] S.R. Niezgoda, D.T. Fullwood, S.R. Kalidindi, Delineation of the space of 2-point correlations in a composite material system, Acta Materialia 56(18) (2008) 5285-5292.

[134] D.T. Fullwood, S.R. Niezgoda, S.R. Kalidindi, Microstructure reconstructions from 2-point statistics using phase-recovery algorithms, Acta Materialia 56(5) (2008) 942-948.

[135] S.R. Niezgoda, Y.C. Yabansu, S.R. Kalidindi, Understanding and Visualizing Microstructure and Microstructure Variance as a Stochastic Process, Acta Materialia 59 (2011) 6387-6400.

[136] C. Suh, A. Rajagopalan, X. Li, K. Rajan, The application of principal component analysis to materials science data, Data Science Journal 1 (2002) 19-26.

[137] S.R. Niezgoda, A.K. Kanjarla, S.R. Kalidindi, Novel microstructure quantification framework for databasing, visualization, and analysis of microstructure data, Integrating Materials and Manufacturing Innovation 2:3 (2013).

[138] B.L. Adams, X. Gao, S.R. Kalidindi, Finite approximations to the second-order properties closure in single phase polycrystals, Acta Materialia 53(13) (2005) 3563-3577.

[139] S.R. Niezgoda, S.R. Kalidindi, Applications of the Phase-Coded Generalized Hough Transform to Feature Detection, Analysis, and Segmentation of Digital Microstructures, Cmc-Computers Materials & Continua 14(2) (2009) 79-97.

[140] S.R. Niezgoda, D.M. Turner, D.T. Fullwood, S.R. Kalidindi, Optimized structure based representative volume element sets reflecting the ensemble-averaged 2-point statistics, Acta Materialia 58(13) (2010) 4432-4445.

[141] D.T. Fullwood, S.R. Kalidindi, S.R. Niezgoda, A. Fast, N. Hampson, Gradient-based microstructure reconstructions from distributions using fast Fourier transforms, Materials Science and Engineering a-Structural Materials Properties Microstructure and Processing 494(1-2) (2008) 68-72.

[142] B. Bochenek, R. Pyrz, Reconstruction of random microstructures: a stochastic optimization problem, Computational Materials Science 31(1-2) (2004) 93-11.

[143] A.P. Roberts, Statistical reconstruction of three-dimensional porous media from two-dimensional images, Physical Review E 56(3) (1997) 3203.

[144] N. Halko, P.-G. Martinsson, Y. Shkolnisky, M. Tygert, An algorithm for the principal component analysis of large data sets, SIAM Journal on Scientific computing 33(5) (2011) 2580-2594.

[145] V. Rokhlin, A. Szlam, M. Tygert, A randomized algorithm for principal component analysis, SIAM Journal on Matrix Analysis and Applications 31(3) (2009) 1100-1124.

[146] I. Jolliffe, Principal component analysis, Wiley Online Library2005.

[147] H.J.C. Berendsen, S. Hayward, Collective protein dynamics in relation to function, Current Opinion in Structural Biology 10(2) (2000) 165-169.

[148] C.C. David, D.J. Jacobs, Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins, in: D.R. Livesay (Ed.), Protein Dynamics, Humana Press2014, pp. 193-226.

[149] A.E. García, Large-amplitude nonlinear motions in proteins, Physical Review Letters 68(17) (1992) 2696-2699.

[150] G.G. Maisuradze, A. Liwo, H.A. Scheraga, Principal Component Analysis for Protein Folding Dynamics, Journal of Molecular Biology 385(1) (2009) 312-329.

[151] X. Dong, D.L. McDowell, S.R. Kalidindi, K.I. Jacob, Dependence of mechanical properties on crystal orientation of semi-crystalline polyethylene structures, Polymer 55(16) (2014) 4248-4257.

[152] C.A. Becker, F. Tavazza, Z.T. Trautt, R.A. Buarque de Macedo, Considerations for choosing and using force fields and interatomic potentials in materials science and engineering, Current Opinion in Solid State and Materials Science 17(6) (2013) 277-283.

[153] C.A. Becker, F. Tavazza, L.E. Levine, Implications of the choice of interatomic potential on calculated planar faults and surface properties in nickel, Philos Mag 91(27) (2011) 3578-3597.

[154] Z.T. Trautt, F. Tavazza, C.A. Becker, Facilitating the selection and creation of accurate interatomic potentials with robust tools and characterization.

[155] A.M. Gokhale, A. Tewari, H. Garmestani, Constraints on microstructural two-point correlation functions, Scripta Materialia 53 (2005) 989-993.

[156] G.B. Olson, C.J. Kuehmann, Materials genomics: From CALPHAD to flight, Scripta Materialia 70 (2014) 25-30.

[157] G.B. Olson, Computational Design of Hierarchically Structured Materials, Science 277(5330) (1997) 1237-1242.

[158] D.L. McDowell, J. Panchal, H.-J. Choi, C. Seepersad, J. Allen, F. Mistree, Integrated design of multiscale, multifunctional materials and products, Butterworth-Heinemann2009.

[159] C. Ward, Materials genome initiative for global competitiveness, 23rd Advanced Aerospace Materials and Processes (AeroMat) Conference and Exposition, Asm, 2012.

[160] J. Holdren, T. Power, G. Tassey, A. Ratcliff, L. Christodoulou, A National strategic plan for advanced manufacturing, US National Science and Technology Council, Washington, DC (2012).

[161] J.C. Meyer, A.K. Geim, M.I. Katsnelson, K.S. Novoselov, T.J. Booth, S. Roth, The structure of suspended graphene sheets, Nature 446(7131) (2007) 60-63.

[162] S.O. Hruszkewycz, M. Allain, M.V. Holt, C.E. Murray, J.R. Holt, P.H. Fuoss, V. Chamard, High-resolution three-dimensional structural microscopy by single-angle Bragg ptychography, Nat Mater 16(2) (2017) 244-251.

[163] A.T. Wicaksono, C.W. Sinclair, M. Militzer, A three-dimensional atomistic kinetic Monte Carlo study of dynamic solute-interface interaction, Modelling and Simulation in Materials Science and Engineering 21(8) (2013) 085010.

[164] M. Berghoff, M. Selzer, B. Nestler, Phase-field simulations at the atomic scale in comparison to molecular dynamics, The Scientific World Journal 2013 (2013).

[165] D. Molnar, R. Mukherjee, A. Choudhury, A. Mora, P. Binkele, M. Selzer, B. Nestler, S. Schmauder, Multiscale simulations on the coarsening of Cu-rich precipitates in α-Fe using kinetic Monte Carlo, molecular dynamics and phase-field simulations, Acta Materialia 60(20) (2012) 6961-6971.

[166] K. Rajan, Combinatorial materials sciences: Experimental strategies for accelerated knowledge discovery, Annu. Rev. Mater. Res. 38 (2008) 299-322.

[167] S. Curtarolo, G.L. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design, Nature materials 12(3) (2013) 191-201.

[168] G. Pilania, P.V. Balachandran, C. Kim, T. Lookman, Finding new perovskite halides via machine learning, Frontiers in Materials 3 (2016) 19.

[169] G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Scientific Reports 6 (2016) 19375.

[170] D.B. Brough, D. Wheeler, J.A. Warren, S.R. Kalidindi, Microstructure-based knowledge systems for capturing process-structure evolution linkages, Current Opinion in Solid State and Materials Science (2016).

[171] D.B. Brough, A. Kannan, B. Haaland, D.G. Bucknall, S.R. Kalidindi, Extraction of Process-Structure Evolution Linkages from X-ray Scattering Measurements using Dimensionality Reduction and Time Series Analysis, Integrating Materials and Manufacturing Innovation (in press) (2017).

[172] S.R. Kalidindi, S.R. Niezgoda, G. Landi, S. Vachhani, T. Fast, A novel framework for building materials knowledge systems, Computers, Materials, & Continua 17(2) (2010) 103-125.

[173] T. Fast, S.R. Niezgoda, S.R. Kalidindi, A new framework for computationally efficient structure–structure evolution linkages to facilitate high-fidelity scale bridging in multi-scale materials models, Acta Materialia 59(2) (2011) 699-707.

[174] D.B. Brough, D. Wheeler, S.R. Kalidindi, Materials knowledge systems in Python - a data science framework for accelerated development of hierarchical materials, Integrating Materials and Manufacturing Innovation (in press) (2017).

[175] P. Steinmetz, Y.C. Yabansu, J. Hötzer, M. Jainta, B. Nestler, S.R. Kalidindi, Analytics for microstructure datasets produced by phase-field simulations, Acta Materialia 103 (2016) 192-203.

[176] J.S. Weaver, A. Khosravani, A. Castillo, S.R. Kalidindi, High throughput exploration of process-property linkages in Al-6061 using instrumented spherical microindentation and microstructurally graded samples, Integrating Materials and Manufacturing Innovation 5(1) (2016) 1-20.

[177] A. Khosravani, A. Cecen, S.R. Kalidindi, Development of High Throughput Assays for Establishing Process-Structure-Property Linkages in Multiphase Polycrystalline Metals: Application to Dual-Phase Steels, Acta Materialia submitted (2016).

[178] A. Gupta, A. Cecen, S. Goyal, A.K. Singh, S.R. Kalidindi, Structure-Property Linkages for Non-Metallic Inclusions/Steel Composite System using a Data Science Approach, Acta Materialia 91 (2015) 239–254.

[179] G. Gottstein, L.S. Shvindlerman, M. Crumbach, L.A. Barrales-Mora, Recent Advances in the Simulation of Recrystallization and Grain Growth, Materials Science Forum, Trans Tech Publ, 2007, pp. 3-12.

[180] A. Khorshidi, A.A. Peterson, Amp: A modular approach to machine learning in atomistic simulations, Computer Physics Communications 207 (2016) 310-324.

[181] V. Botu, R. Batra, J. Chapman, R. Ramprasad, Machine Learning Force Fields: Construction, Validation, and Outlook, The Journal of Physical Chemistry C 121(1) (2017) 511-522.

[182] J. Schiøtz, K.W. Jacobsen, A Maximum in the Strength of Nanocrystalline Copper, Science 301(5638) (2003) 1357-1359.

[183] S.V. Kalinin, Scanning Probe Microscopy in US Department of Energy Nanoscale Science Research Centers: Status, Perspectives, and Opportunities, Advanced Functional Materials 23(20) (2013) 2468-2476.

[184] S.J. Billinge, M.G. Kanatzidis, Beyond crystallography: the study of disorder, nanocrystallinity and crystallographically challenged materials with pair distribution functions, Chemical communications (7) (2004) 749-760.

[185] T. Proffen, S.J.L. Billinge, T. Egami, D. Louca, Structural analysis of complex materials using the atomic pair distribution function — a practical guide, Zeitschrift für Kristallographie - Crystalline Materials, 2003, p. 132.

[186] S.V. Kalinin, B.G. Sumpter, R.K. Archibald, Big-deep-smart data in imaging for guiding materials design, Nat Mater 14(10) (2015) 973-980.

[187] S. Jesse, B.J. Rodriguez, S. Choudhury, A.P. Baddorf, I. Vrejoiu, D. Hesse, M. Alexe, E.A. Eliseev, A.N. Morozovska, J. Zhang, L.-Q. Chen, S.V. Kalinin, Direct imaging of the spatial and energy distribution of nucleation centres in ferroelectric materials, Nat Mater 7(3) (2008) 209-215.

[188] S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular-Dynamics, J Comput Phys 117 (1995).

[189] Y. Mishin, D. Farkas, M.J. Mehl, D.A. Papaconstantopoulos, Interatomic potentials for monoatomic metals from experimental data and ab initio calculations, Phys Rev B 59 (1999).

[190] L. Priester, Grain Boundary Order/Disorder and Energy, Grain Boundaries: From Theory to Engineering, Springer Netherlands, Dordrecht, 2013, pp. 93-132.

[191] M.S. Daw, M.I. Baskes, Semiempirical, quantum mechanical calculation of hydrogen embrittlement in metals, Physical review letters 50(17) (1983) 1285.

[192] M.S. Daw, M.I. Baskes, Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals, Physical Review B 29(12) (1984) 6443.